University of São Paulo
"Luiz de Queiroz" College of Agriculture

Statistical models to study of reproductive biotechnology in cattle

Andreza Jardelino Koeller

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2023

Andreza Jardelino Koeller
Bachelor in Statistics

Statistical models to study of reproductive biotechnology in cattle

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof^a. Dr^a. **CLARICE GARCIA BORGES DEMÉTRIO**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Koeller, Andreza Jardelino

Statistical models to study of reproductive biotechnology in cattle / Andreza Jardelino Koeller. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2023 .

74 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Dados binários 2. Modelos combinados 3. Modelos lineares generalizados mistos 4. Machine learning 5. Superdispersão 6. Classificação de prenhez 7. Random forest 8. Embrião viável. I. Título.

To my husband Brian Koeller, this achievement I dedicate to you.

ACKNOWLEDGEMENTS

I would like to thank God, for his infinite compassion in my life, for the strength and coherence in my attitudes, for the achievement of my dreams.

My husband Brian, for all the emotional and intellectual support, for without your assistance I would not have made this far. I know these have been hard times, but at every moment you made me persevere, keeping alive in me the hope of better days. I love you and thank God for having you in my life.

To my adviser, professor Clarice, who agreed to guide me and contributed to my professional development with your immense knowledge, for the patience and wisdom in handling the difficulties we met along the way. Thank you for all you have done for me.

To Daniela Demétrio, for providing all the data needed for this work, as well as contributing with all the knowledge necessary for the development and support of this study.

This work was financially supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Finally, I would like to thank all of those, who have contributed either directly or indirectly for the realization of this work, my most sincere gratitude.

CONTENTS

RESUMO

Modelos estatísticos para o estudo de biotecnologias reprodutivas em bovinos

O advento das técnicas de inseminação artificial e fertilização *in vitro* possibilitaram que o campo de melhoramento animal obtivesse um avanço nos resultados de prenhez. Os dados utilizados nesse trabalho são referentes a transferência e viabilidade embrionária, onde a natureza da variável de interesse são amostras binárias e de proporção, respectivamente. Nesse contexto, o objetivo deste trabalho é desenvolver modelos capazes de acomodar esses tipos de dados, e com isso avaliar as possíveis influências para cada um dos interesses, fomentando o conhecimento no que tange a área de estatística, bem como a de melhoramento animal. Para o desenvolvimento deste trabalho foram utilizados modelos lineares generalizados mistos para avaliar os dados binários superdispersos, a fim de identificar quais os fatores que influenciavam a uma transferência embrionária de sucesso. Um outro objetivo foi verificar as condições que resultam em uma elevada taxa de viabilidade embrionária, para isso, foram propostos os modelos combinados para acomodar os dados de proporção. Para finalizar o trabalho foi proposta uma comparação entre de diferentes metodologias, as quais utilizaram os dados binários a fim de verificar a performance entre modelos estatísticos com aqueles propostos para o aprendizado de máquina.

Palavras-chave: Dados binários, Modelos combinados, Modelos lineares generalizados mistos, Machine learning, Superdispersão, Classificação de prenhez, Random forest, Embrião viável.

ABSTRACT

Statistical models to study of reproductive biotechnology in cattle

The advent of artificial insemination and *in vitro* fertilization have made it possible for the field of animal breeding to gain sizeable advances in pregnancy outcomes. The data used in this work relates to embryo transfer and viability, where the nature of the variable under study are, respectively, binary and proportion samples. In this context, the goal of this work is to develop models capable of accommodating these kinds of data, and with data evaluate the possible influences for each of the interests, advancing knowledge in the field of statistics, as well as animal breeding. For the development of this work generalized linear mixed-effects models to evaluate the overdispersed binary data, with the objective of identifying which factors influenced a successful embryo transfer. Another goal was to verify which conditions lead to a high embryo viability rate, for that, combined models were proposed as a solution capable of accommodating the proportion data. Finally, in the last chapter, we proposed a comparison between different methodologies, which used the binary data with the objective of verifying the performance between statistical models with those proposed for machine learning.

Keywords: Binary data, Combined models, Generalized linear mixed models, Machine learning, Overdispersion, Pregnancy classification, Random forest, Viable embryo.

1 INTRODUCTION

The reproductive efficiency in cattle ensures greater productivity and profitability for the producer. This advance in the supply of products from the livestock is due to the development of reproductive techniques, and animal breeding.

The techniques of production and embryo transfer have contributed to the reproductive performance since it facilitates acceleration in the process of multiplication of animals of higher zootechnical value. According with Barros et al. (1995), a bovine female shows approximately a 21 day estrous cycle, that is, under ideal conditions, without the use of reproductive biotechniques, it would only be possible to produce one calf per year. Thus, superovulation is used through the application of exogenous hormones to later perform the artificial insemination (AI), and transfer them to the receiving females. One of the major obstacles to employment in the larger volume of these techniques is the great variability of the response.

In this context, researchers are interested in the identification of animals with greater genetic potential, as well as in the main factors that influence the multiplication process.

Studies show that there is growing concern about statistical modeling, since one must consider the nature of the variable of interest, whether continuous or discrete. Among the great variety of models, Nelder e Wedderburn (1972) proposed the Generalized Linear Models (GLMs) that associate to the response variable to a probability distribution since this can be written in the form of the exponential family.

By combining GLMs with random effects in the linear predictor, we have the Mixed Generalized Linear Models (GLMMs). Developed by Lindsay (1986); Breslow e Clayton (1993), the theory allows describing several sources of variation, caused by the inclusion of unobserved latent variables, which are responsible for the excess variability. In addition, it allows accommodating dependency structures between observations. However, when constructing models based on this methodology, some constraints are imposed, such as correctly specifying the response variable, defining the components of the model, and including random effects and covariance structures appropriately.

In practice, for data that involve response variable in the form of proportions or counts, the inclusion of random effects will not always be sufficient, since they may be overdispersed, that is, the variability of the data is greater than expected by the specified model. It is necessary in this case to use models whose structure makes it possible to add an extra variation. The implication of data adjustments in which overdispersion is not considered is the overestimation of the deviance associated with the terms of the model and the underestimation of the standard errors of the parameter estimates, which will consequently promote erroneous interpretations of the significance of the effects, inducing in this case, the incorrect selection of models (Hinde e Demétrio, 1998; Moral et al., 2017).

In this context, in order to accommodate this extra variability, Hinde e Demétrio (1998) incorporated a dispersion parameter into the variance function. Later, Molenberghs et al. (2007, 2010) would combine the GLMMs with those models that consider the phenomenon of overdispersion.

There also exist in the literature other techniques which have been found to successful model binary data, the machine learning algorithms, which have found a vast array of appli-

cations in the discipline which came to be known as data science. This approach has found broad acceptance in industry, and powers much of the analytical capabilities behind common technologies such as search engines (Mahesh, 2020).

In the existing literature, no applications of these methodologies to bovine reproduction data were found. Thus, the objectives of the work were to study and propose models, which could be used to solve problems related to production and embryo transfer data, such as, overdispersion and correlation structures. In Chapter 2, the case study was related to the embryo transfer, being proposed the Bernoulli distribution, with the response variable corresponding to the receiving cow's pregnancy diagnosis, being of the binary type, and analyzed via hierarchical models. In Chapter 3, the embryo production data set was used, with the response variable rate of embryonic viability, which was modeled considering the binomial distribution using the methodology of the combined models. In Chapter 4, we proposed a comparison between different methodologies, such as classical statistics and those applied in data science, such as machine learning algorithms. Finally, in Chapter 5 we have the final considerations obtained in this study and future perspectives to the research.

References

Barros, C., Figueiredo, R., e Pinheiro, O. (1995). Estro, ovulação e dinâmica folicular em zebuínos. *Revista Brasileira de Reprodução Animal*, 19(1-2):9–22.

Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Hinde, J. e Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Lindsay, B. G. (1986). Exponential family mixture models with least-squares estimators. *The Annals of Statistics*, pages 124–137.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, 25(3):325–347.

Moral, R. A., Hinde, J., e Demétrio, C. G. (2017). Half-normal plots and overdispersed models in r: The hnp package. *Journal Statistical Software*, 81(10):1–23.

Nelder, J. e Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135:370–384.

2 HIERARCHICAL MODEL APPLIED TO EMBRYO TRANSFER DATA

Abstract

In 2021, there were more than 1.5 million embryo transfers (ET) recorded worldwide. The main goal of any embryo program is to obtain a pregnancy and consequently a live calf and there are many variables that can affect the results. Embryo transfer studies are usually not planned experiments but observational data which are very unbalanced and heterogeneous. The pregnancy outcome is a binary response of either a success or a failure. Data cases of this nature often present a typical case of variance much greater than that specified by the model, characterized as overdispersion. The limitations in the statistical analyses of binary data were solved with the emergence of Generalized Linear Models, which enriched expressively, with respect to flexibility regarding classification, and choice of models for the response variable. A parametric alternative in these situations is to adopt the methodology of the Mixed Generalized Linear Models, which allows to include one or more random effect terms in the linear predictor. In this study, the Bernoulli-logistic-normal model was applied to embryo transfer data, with pregnancy outcome being the interest of this study. The results show that the approach proposed in relation to conventional models was able to capture the Bernoulli model's lack of fit and possible sources of dispersion present in the data set. In addition, it helped in understanding the factors that may influence ET pregnancy outcomes.

keywords: Bernoulli Distribution; Generalized Linear Mixed Model; Pregnancy Classification.

2.1 Introduction

Embryo production has proven to be a powerful technology for bovine genetic improvement, primarily to propagate the genes of females with superior genetic values and lineage. In 2021, there were more than 1.5 million embryo transfers (ET) recorded worldwide (Viana, 2022). The main goal of any embryo program is to obtain a pregnancy and consequently a live calf. There are many variables that can possibly affect pregnancy outcomes such as embryo stage of development, embryo quality, embryo transfer practitioner, season, type of semen (conventional or sex sorted), donor dam and sires, recipient status, and possibly genomic values of the donor, sire and recipients such as fertility index (FI), daughter pregnancy rate (DPR) and total performance index (TPI) (Hansen, 2020).

Embryo transfer studies are usually not planned experiments but observational data which are very unbalanced and heterogeneous. An important step of the data analysis is the data cleaning which consists in the process of identifying, formatting, parsing and fixing or removing data from a dataset to improve data quality by eliminating wrongful or inaccurate information (Wickham, 2014). It is the process of structuring data in such a way that makes it easy to inspect, visualize or analyze it. The data is checked for completeness, and any abnormal data is considered unknown. An important point to be highlighted is that the statistician should always work with the data collector to find mistakes in the data set and also to understand the factors that have relevance in the analysis.

The pregnancy outcome is a binary response of either a success (1) or a failure (0). A critical analysis of the pregnancy data depends on an adequate statistical analysis for better interpretation of the data, taking into account the variables that can possibly affect pregnancy. The use of the traditional statistical analysis, ANOVA, does not apply to the study of the pregnancy data, since the data usually do not follow a normal distribution.

A Bernoulli regression model, a particular case of the generalized linear models (GLMs) (McCullagh e Nelder, 1989), provides a standard framework for the analysis of binary data. However, a common problem is the potential of overdispersion that occurs when the data display more variability than is predicted by the variance-mean relationship for the assumed model. Also, in this type of study it is very common to have a mixture of crossed and hierarchical data, given that different sires can be mated with different dams and/or some common ones. This will bring different types of correlation between the embryos, for example, embryos generated with the semen of a bull with oocytes of a dam will be more correlated than embryos generated with the semen of a bull with oocytes of different dams. The possible correlation between measurements resulted from the clustering is often accommodated through the inclusion of subject-specific, random effects (Molenberghs et al., 2007, 2010, 2017).

The structures of correlations between individuals and/or variability greater than that specified by the assumed distribution for the variable of interest, should be considered in the model, being necessary the addition of one or more random effects in the linear predictor, characterizing the Mixed Generalized Linear Models.

If overdispersion and/or correlation are not taken into account all model selection criteria would generally be expected to perform poorly and a model with too many parameters is likely to be selected. The estimates of the parameters of the model and its standard errors will be incorrect and we may incorrectly assess significance of individual parameters.

The aim of this paper is to propose models that best fit ET observational data with a binary response, and describe ways to analyze these types of data with the goal of identifying which factors influence the ET pregnancy outcome.

In this work we review and compare methods for analyzing binary data with particular focus on potential applications in agricultural research. In subsection 2.2.1 provides a motivation data set. In subsection 3.2.3 and subsection 2.2.3 presents some models used for the analysis of binary data, and discusses model selection and diagnostics. The motivation data set is analyzed and discussed in section 2.3. Some general considerations are presented in section 2.4. The scripts developed in the software **R** (R Core Team, 2018) are presented in the Appendix.

2.2 Material and Method

2.2.1 Case Study: Embryo Transfer Data

To better understand the effects that influence ET pregnancy outcomes, data was collected from 2015 to 2019 at RuAnn and Maddox Dairy Farms in Riverdale, California, USA, that milks approximately 4500 lactating Holstein cows and have been doing ET to multiply their best females since 1982. After cleaning up the initial data for typing errors that could influence analysis results, a total of 5108 fresh or frozen in vivo derived embryos (IVD) were transferred

to heifers or lactating cow recipients. It consisted of 4070 recipients, 386 donor cows, 176 sires, 940 different donor-sire combinations. The embryos were transferred by different experienced practitioners (TECH: T1 (2659), T2 (730), T3 (650), T4 (786), T5 (72), T6 (202)). The response variable was Pregnancy Diagnosis (PD: 0 - Open (1955), 1 - Pregnant (3153)).

The dates of ET were recorded and divided by seasons of the year (Season: 1 - Summer (1158; July, August, September), 2 - Autumn (1240; October, November, December), 3 - Winter (1411; January, February, March), 4 - Spring (1299; April, May, June),). The embryos transferred were classified according to type (TEEF2: 1 - fresh grade 1, 2 - fresh grade 2, 3 - frozen-thawed grade 1) and according to the stage of development (embryo_stage: 4 - morula, 5 - early blastocyst, 6 – blastocyst), as described by the International Embryo Technology Society (Wright, 2001).

The donor cows were individually recorded and divided by status (TYPE: 1 – dry cow (2703), 2 – heifer (323), 3 – lactating cows (2082)), total performance index (TPI_D) and information of fertility index (FI_D) were recorded.

All donor sires were recorded and divided by type of semen (SEXED: 1 - not sexed (4563) or 2 – sexed (545)) and information of fertility index (FI_S). The recipients were individually recorded and divided by status (VH_LC: 1 – lactating cow (1314), 2 – heifer (3794)) and information of fertility index (FI_R), total performance index (TPI_R) and daughter pregnancy rate (DPR_R) was recorded. The side of corpus luteum (CL) on ET day (CL_side: 1 – left (2137) or 2 - right(2971)) and the number of days after estrus (DFE: 6, 7, 8, 9 days) were also recorded.

Initially, we carried out an exploratory data analysis (Tukey et al., 1977), by using several graphical representations, to gain further insight into the available data and identify any hidden patterns, trends, and relationships in data before developing statistical methods for hypothesis testing (confirmatory data analysis).

All the transferred embryos were IVD. The recipients were observed daily for spontaneous estrus and received an embryo 6 to 9 days after estrus in the uterine horn ipsilateral to the CL. Pregnancy diagnosis was conducted by rectal palpation 26 to 40 days after ET (33 to 40 days after estrus).

2.2.2 Statistical Approach

The pregnancy outcome is a binary response of either a success (1) or a failure (0). A Bernoulli regression model provides a standard framework for the analysis of binary data and is a particular case of the generalized linear models (McCullagh e Nelder, 1989)

Generalized linear models

Generalized Linear Models involve three components (Nelder e Wedderburn, 1972):

i) A *random component*, represented by the independent random variables $Y_i, i = 1, \cdots, n,$, which have the same distribution belonging to the exponential family in canonical form, given by

$$f_Y(y_i; \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \tag{2.1}$$

where $\phi > 0$ is a constant dispersion parameter, θ_i is the canonical parameter, $b(.)$ and $c(.,.)$ are known functions. The mean and variance are, respectively, $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$ and $\mathrm{Var}(Y_i) = \phi b''(\theta_i) = \phi \mathrm{V}(\mu_i)$, with $\mathrm{V}(\mu_i) = d\mu_i/d\theta_i$ called the variance function.

ii) A *systematic component* represented by the explanatory variables, $x_1, \cdots, x_n$, included in the form of a linear sum of the effects as a linear predictor

$$\eta_i = \sum_{j=1}^{P} x_{ij}\beta_j = x_i^T\beta \qquad i = 1, \ldots, n,$$

or,

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

where η is the vector of linear predictors, X the design matrix of the model $(n \times p)$ and β is the vector $(p \times 1)$ of parameters to be estimated.

iii) A *link function* relating the mean to the linear predictor,

$$g(\mu_i) = \eta_i,$$

where $g(\cdot)$ is a real function, monotonous and differentiable (McCullagh e Nelder, 1989).

Therefore, to fit a GLM it is necessary to choose a distribution for the response variable, the matrix to represent the linear predictor of the model, and a link function.

The estimation of the parameter vector β is by maximum likelihood, and based on a Fisher scoring algorithm (Nelder e Wedderburn, 1972) results in an iteratively weighted least squares algorithm and at convergence

$$\hat{\beta} = (\mathrm{X}^T\mathrm{W}\mathrm{X})^{-1}\mathrm{X}^T\mathrm{W}\mathrm{z},$$

where X is the design matrix of the model, W is a diagonal matrix with elements $W_i = \frac{1}{V(\mu_i)[g'(\mu_i)]^2}$, $g'(\mu_i) = \frac{dg(\mu_i)}{d\mu_i}$ and z is the adjusted response variable with $z_i = \eta_i + g'(\mu_i)(y_i - \mu_i)$. The estimates minimize the deviance function given by

$$\begin{aligned} S_p &= 2(\hat{\ell}_n - \hat{\ell}_p) = \phi^{-1}D_p \\ &= 2\phi^{-1}\sum_{i=1}^{n}[y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)], \end{aligned} \tag{2.2}$$

where S_p is the scaled deviance, D_p is the deviance; $\hat{\ell}_n$ and $\hat{\ell}_p$ are the maximum of the logarithm of the likelihood function for the saturated and under study models, respectively and $\hat{\theta}_i$ and $\tilde{\theta}_i$ are the maximum likelihood estimates of the canonical parameter, under the saturated and reduced models. It is a goodness of fit measure of the distance between the observed and fitted values in units of log-likelihood.

Another goodness of fit measure is generalized Pearson's statistic X^2 defined by

$$X_p^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \tag{2.3}$$

where $V(\hat{\mu}_i)$ is the estimated variance function for the distribution under study (McCullagh e Nelder, 1989).

When the dispersion parameter ϕ is known, the scaled deviance and the statistic X_p^2/ϕ, follow, asymptotically, a χ^2 distribution with $(n-p)$ degrees of freedom. Furthermore, as the sample size increases, the generalized Pearson statistic converges more quickly to the reference distribution. However, in practice, in general, deviance and X^2 are much greater than the one specified by the model and can be evidence of overdispersion. Using an inadequate model that does not account for overdispersion may lead to make overly precise inferences and predictions, as certainly standard errors will be incorrect and may be seriously underestimated. When ϕ is unknown there is no formal test for checking the adequacy of the model, which can be investigated using residuals analysis (Jørgensen, 2002).

2.2.3 Statistical Approach: Data Modelling

Binary data

Let's assume that the random variables Y_i represent an individual binary outcome, pregnant or open. The basic probability model for the random variable Y_i is the Bernoulli distribution, Bern(π_i); it has a probability mass function given by

$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1, \ldots, n,$$

where π_i is the probability of a successful pregnancy and $y_i = 0$ (open) or $y_i = 1$ (pregnant). The mean and variance are, respectively, $\mathrm{E}(Y_i) = \pi_i$ $\mathrm{Var}(Y_i) = \pi_i(1-\pi_i)$. Then, a generalized linear model allows us to model the probability of success in terms of the exploratory variables x_i through

$$\eta_i = g(\pi_i) = x_i^T\beta, \quad i = 1, \ldots, n,$$

where g is some suitable link function and β is a vector of p unknown parameters. For the canonical logit link function

$$\eta_i = \log\frac{\pi_i}{1-\pi_i}.$$

For Bernoulli data there is no overall goodness-of-fit test and assessments of model adequacy can only be made by fitting extended models or by grouping data (Hinde e Demétrio, 1998).

Proportion data

If we have replicate binary variables for each distinct covariate, the individual binary responses can be grouped to give number of successes out of number of trials and treated as binomial. Let's assume that the random variables Y_i represent numbers of pregnancies out of samples of size m_i, $i = 1, \ldots, n$. The standard probability model for the random variable Y_i is the Binomial distribution, $Y_i \sim \mathrm{Bin}(m_i, \pi_i)$, $\mathrm{B}(m_i, \pi_i)$; it has a probability mass function given by

$$P(Y_i = y_i) = \binom{m_i}{y_i}\pi_i^{y_i}(m_i-\pi_i)^{m_i-y_i}, i = 1, \ldots, n,$$

where π_i is the probability of a successful pregnancy and $y_i = 0, 1, \ldots, m_i$. The mean and variance are, respectively, $\mathrm{E}(Y_i) = m_i\pi_i$ and $\mathrm{Var}(Y_i) = m_i\pi_i(1-\pi_i)$. The probability of success can be modelled in terms of the exploratory variables x_i through

$$\eta_i = \log \frac{\pi_i}{1-\pi_i} = x_i^T\beta, \quad i = 1, \ldots, n.$$

For binomial distribution, $\phi = 1$, and subject to certain asymptotic conditions, for a well fitting model we would expect, in an exploratory way, that the deviance, $S_P = D_p$, and the statistic X_p^2 to be approximately equal to the residual degrees of freedom.

Overdispersion for binary and proportion data

Overdispersion can arise in various ways, typically through some failure of the basic model assumptions. The assumptions for the binomial model are independence of observations and constant probability of success. If one or both of these assumptions failure the variance of the data will be greater than the variance expected using the binomial model, $\mathrm{Var}(Y_i) > m_i\pi_i(1-\pi_i)$, resulting in overdispersion.

It can be caused by the variability of the experimental material, poor specification of the linear predictor, excess of zeros, omitted variables in the linear predictor, correlation between individual responses and cluster sampling, that makes the probability of success not constant for all observations (Hinde e Demétrio, 1998). In general, it is difficult to infer the precise cause, or underlying process, leading to overdispersion.

Many different specific models can arise from alternative possible mechanisms for the underlying process. The simplest way to accommodate overdispersion is to assume some more general form for the variance function, possibly including additional parameters, leading to the quasi-binomial model.

Assuming now that the pregnancies are not happening independently or are at some varying underlying rate (e.g. differences in fertility of the females), contributing additional variability to the recorded observations, e.g. a two-stage model could be assumed for the response, that is, the response variable follows a binomial distribution, $Y_i|P_i \sim \mathrm{B}(m_i; P_i)$, and the parameter itself has a beta distribution $P_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$ resulting in a beta-binomial model. The mean and variance of Y_i are, respectively, $\mathrm{E}(Y_i) = m_i\pi_i$ and $\mathrm{Var}(Y_i) = m_i\pi_i(1-\pi_i)[1+\phi(m_i-1)]$ (Hinde e Demétrio, 1998).

Generalized linear mixed model

Since we may think that there is a combination of many unexplained sources affecting the success of a pregnancy we can include a normal random effect at observation level, $U_i \sim N(0, \sigma_u^2)$, in the linear predictor,

$$\eta_i = x_i^T\beta + u_i, \quad i = 1, \ldots, n, \tag{2.4}$$

where U_i is a random effect with variance σ_u^2 and β are the fixed effects, resulting the binomial-normal model, an example of a generalized linear mixed model (GLMM), allowing to get a measure of intraclass correlation.

Another reason for extending the Bernoulli/binomial model is because of the occurrence of a hierarchical structure in the data caused by a clustering resulted from repeatedly measuring the outcome on the same experimental unit (e.g. embryos obtained with semen from the same sire or oocytes from the same cow). The possible correlation between measurements for the same individual is often accommodated through the inclusion of subject-specific, random effects (Verbeke e Molenberghs, 2000).

Let $Y|u$ be a vector of n independent responses conditional on $U = u$ with a distribution that belongs to the exponential family expressed by Equation 2.1. The linear predictor now is

$$\eta = X\beta + Zu,$$

where β is a vector of unknown parameters, u is a vector of unobservable realizations of a random variable U with $U \sim \mathrm{N}(0, G)$, X and Z are design matrices for the fixed and random effects.

For the particular case of a binary response variable, we have the Bernoulli-logistic-normal model defined by

$$\begin{aligned} Y_{ij}|u_i &\sim \text{Bernoulli}(\pi_{ij}), \\ \pi_{ij} &= \frac{\exp(x_i^T\beta + z_i^T u)}{1 + \exp(x_i^T\beta + z_i^T u)}, \\ U &\sim \mathrm{N}(0, G), \end{aligned} \tag{2.5}$$

where Y_{ij}, is the j-th measurement ($j = 1, \cdots, n_i$) of the i-th cluster ($i = 1, \cdots, N$).

Combined model

Additionally, overdispersion and correlation between observations may occur simultaneously, and models accommodating them at once are less than common. Molenberghs et al. (2007, 2010) propose a generalized linear model, accommodating overdispersion and clustering through two separate sets of random effects.

Considering the particular case of proportion data, we assume n independent responses Y_{ij} conditional on the random effects θ_{ij} and u_i, with *j-th* cluster measurement ($j = 1, \cdots, n_i$) of the *i-th* observation ($i = 1, \cdots, N$), with a binomial distribution and a logistic link. Then

$$\begin{aligned} Y_i|b_i, \theta_{ij} &\sim \text{Binomial}(\theta_{ij} k_{ij}), \\ k_{ij} &= \frac{\exp(x_i^T\beta + z_i^T u_i)}{1 + \exp(x_i^T\beta + z_i^T u_i)}, \\ U_i &\sim \mathrm{N}(0, G) \quad \text{and} \quad \theta_{ij} = \text{Beta}(\alpha_{ij}, \beta_{ij}). \end{aligned}$$

where Y_{ij}, is the j-th measurement ($j = 1, \cdots, n_i$) of the i-th cluster ($i = 1, \cdots, N$), β is a vector of fixed effect parameters, x_i^T and z_i^T are the *i-th* lines of the fixed and effects design matrices, u is the vector of random effects, where $u \sim \mathrm{N}(0, G)$, G is the variance-covariance matrix (McCulloch e Searle, 2001). The resulting model is a beta-binomial-normal model (Molenberghs et al., 2010).

Estimation of the parameters

The estimation of the fixed parameter vector β, and the components of variance of the matrix G, is done by maximizing the marginal likelihood function, which is obtained by

integrating the likelihood function with respect to the random effects (Molenberghs et al., 2007, 2010).

All analyses were implemented in the software R (R Core Team, 2018). The parameter estimates were obtained by maximum likelihood, using the Gauss-Hermite adaptive quadrature algorithm, implemented in the glmer function of the package lme4 (Bates et al., 2015).

Goodness of fit and diagnostics

After fitting a model to a data set, in addition to the global goodness-of-fit, it is useful to use some diagnostic plots to detect specific aspects of possible model failure. Typical plots are the dispersion plots of fitted versus observed values, Pearson or deviance residuals versus fitted values among others. To check that the residuals are consistent with the variation implied by the model, an approach is to use the half normal plot with simulated envelope which is implemented in the hnp package (de Andrade Moral et al., 2017).

Model selection - Inference for components of variance

To test random terms, useful asymptotic test is the likelihood-ratio test (LR) that are based on comparing the values of likelihood functions of two nested models, having the same set of fixed-effect parameters, but different sets of covariance parameters. The likelihood ratio statistics is given by

$$LR = -2[\text{logLik(reduced model) - logLik(complete model)}].$$

where logLik is the logarithm of the likelihood function.

When there is no parameter on the boundary of the parametric space, $LR \sim \chi^2_\nu$, where ν is the difference in number of degrees of freedom between the two models. LR has a distribution that is a mixture of χ^2's when there are parameters on the boundary of the parametric space (Self e Liang, 1987). In the case of component of variance models with independence between the random effects as in our case the mixture of χ^2's is given by

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m. \tag{2.6}$$

Akaike (AIC) and Bayesian (BIC) Information Criteria are used when the two models being compared are non-nested AIC (AKAIKE, 1973; Schwarz, 1978),

$$\text{AIC} = -2\text{logLik} + 2p \quad \text{and} \quad \text{BIC} = -2\text{logLik} + \log(n)p.$$

where p is the number of fitted parameters and n is the number of observations.

Model selection - Inference for fixed effects

After choosing a model for random terms, to test for fixed effects we can use the likelihood-ratio test (LR) that are based on comparing the values of likelihood functions of two nested models, having the same set of random-effect parameters, but different sets of fixed parameters.

Akaike (AIC) and Bayesian (BIC) Information Criteria can be used when the two models being compared are non-nested (AKAIKE, 1973; Schwarz, 1978).

To help with the selection of a model the drop1 function from stats package computes all the single terms in the scope argument that can be added to or dropped from the model, fit those models and compute a table of the changes in fit. It gives a comparison of models based on the AIC criterion and when using the option test="F" adds a "type II ANOVA" (using Anova function from the car package). The hierarchy is respected when considering terms to be added or dropped: all main effects contained in a second-order interaction must remain, and so on.

2.3 Results and Discussion

2.3.1 Exploratory analysis

In order to understand the generating process of the data and to visualize possible interactions between factors, before fitting models some descriptive analyses are presented. An initial examination of the dispersion plots of pregnancy data (0/1) versus the fertility indices (FI) of the donor cows, and recipient cows with a superimposed smooth curve Figure 2.1 shows that there is some suggestion that only the extreme values of FI for the recipient cows can affect the probability of success with a small difference for Heifer and lactating cow. The plot is constructed using a simple loess, local polynomial, smooth. However, these plots do not take into account the possible correlation between FI and other covariates.

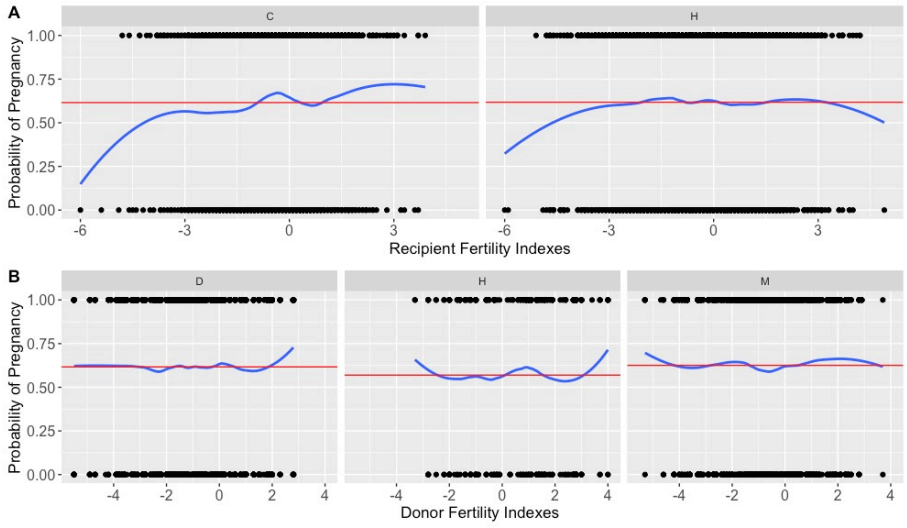

Figure 2.1. Embryo Transfer Data: Pregnancy data (0/1) versus fertility index of the (A) recipient (lactating cows and heifers) and (B) donor (dry cow, heifer, lactating cow) with superimposed smoother (–) and pregnancy estimated probability (–)

The plot of average pregnancy rates versus days after estrus, by embryo stage in Figure 2.2A, shows that ET 7 and 8 after estrus resulted in better pregnancy rates for the three levels of the embryo stage, being better for blastocyst and worse for morula, while for on days 6 and 9 after estrus it is not so clear. The plot of average pregnancy rates versus days after estrus, by type of embryo in Figure 2.2B, shows that embryo transfer on days 7 and 8 after

estrus resulted in better pregnancy rates for the three levels of the type of embryo, being better for fresh grade 1 and worse for fresh grade 2. No evidence of interaction between these factors is seen in both plots.

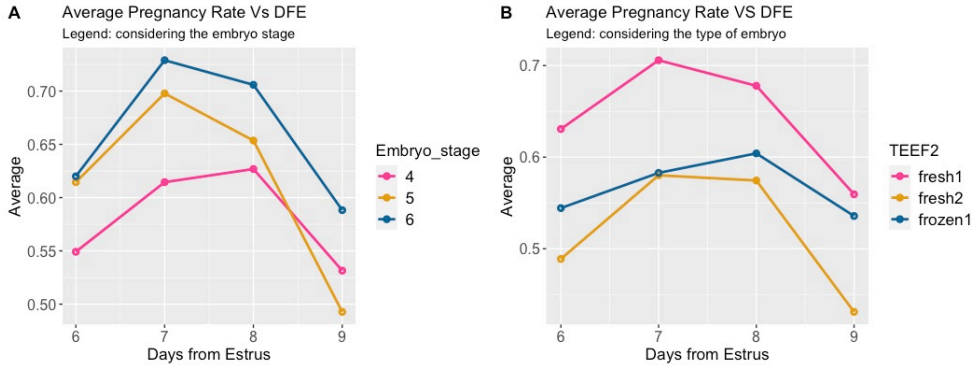

Figure 2.2. Embryo Transfer Data: A) Average pregnancy rates versus days after estrus, by embryo stage; B) Average pregnancy rates versus days after estrus, by type of embryo transfer

Bar plots for observed average pregnancy success rates for each level of the categorical covariates (CL_side, TYPE, TEEF2, embryo_stage and DFE) are presented in Figure 2.3.

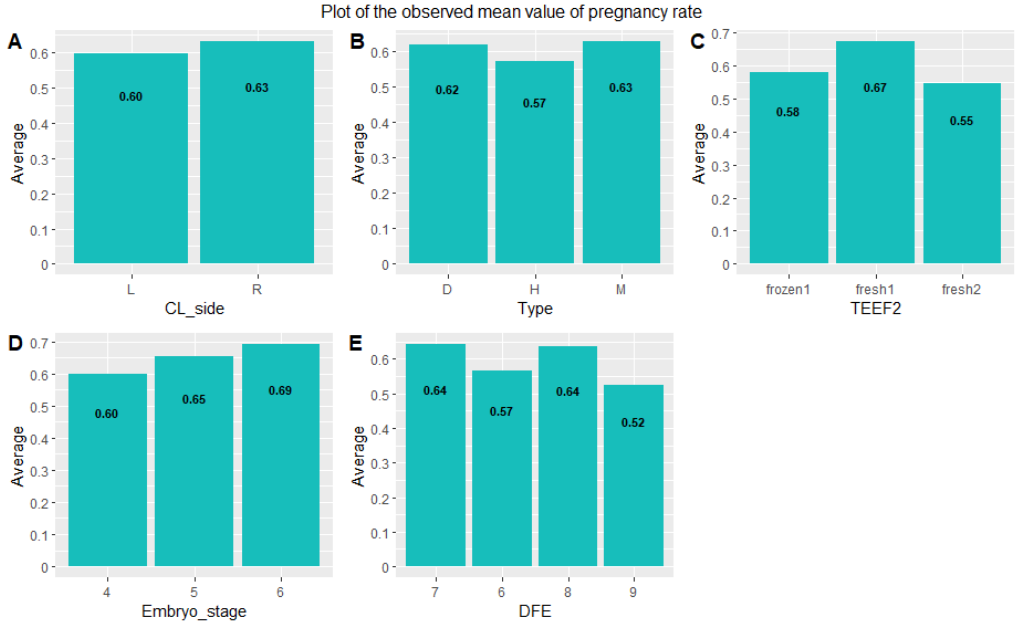

Figure 2.3. Embryo Transfer Data: Observed average pregnancy success rates for each level of the categorical covariates

A boxplot of the frequency of the numbers of pregnant cows versus seasons of the year, by pregnancy diagnosis (Figure 2.4A) and frequency of the number of pregnant cows versus seasons of the year, by type of semen (Figure 2.4B), showed evidence of no differences in the results for the periods of the year in which the transfers were made.

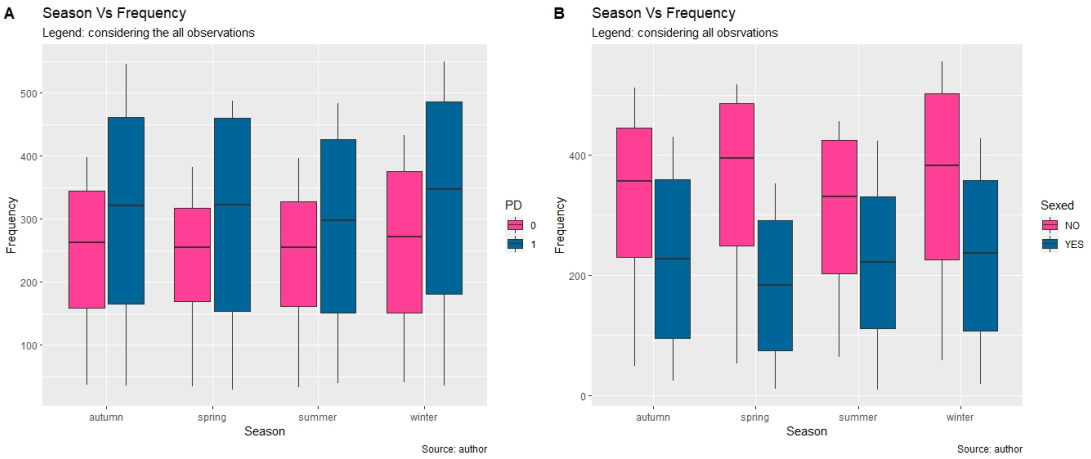

Figure 2.4. Embryo Transfer Data: A) Frequency of numbers of pregnant cows versus seasons of the year, by pregnancy diagnosis (PD); B) Frequency of the number of pregnant cows versus seasons of the year, by type of semen

Simple and partial correlation coefficients between the FI, TPI and DPR of donor and recipients are presented in Figure 2.5A suggest high association between FI and DPR, moderate association between TPI and FI and DPI and FI, and very small association for those variables between donor and recipient cows, as expected.

We have in Figure 2.5B approximate densities for the values of these fertility indexes for donor and cows, showing a symmetric distribution. After the measure of the degree of association between the covariates, it is observed in Figure 2.5C that there was no dependence relationship between the indexes, this result was corroborated by Figure 2.5D, since there is no evidence of a linear relationship between them. Positive high correlations between FI, DPR and TPI to both recipient and donor cows are expected since DPR is part of the FI formula and FI is included in the TPI formula (H.A, 2017).

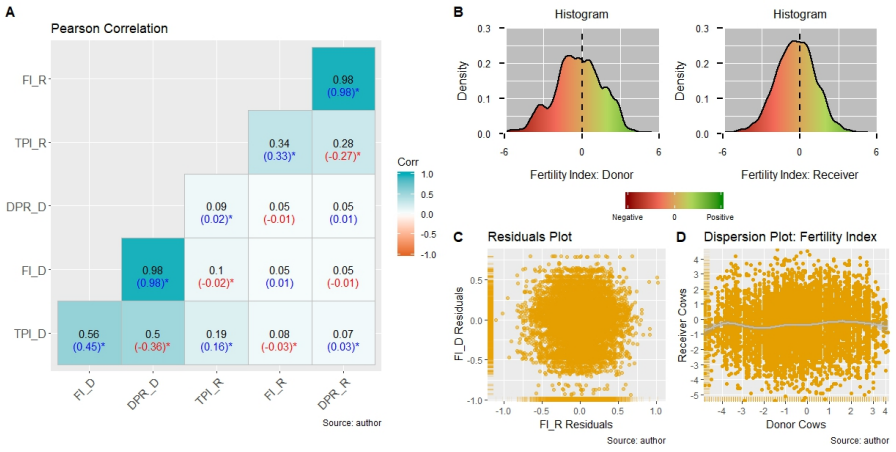

Figure 2.5. Embryo Transfer Data: A) Simple (black) and partial correlations (red - negative associations; blue - positive associations); B) Densities of the fertility indexes of the donor and recipient cows, respectively; C) Residual plot between donor cow and recipient cow; D) Dispersion plot between the fertility indexes of the recipient cow versus donor cow, and a linear regression line based on the Gaussian model (gray). Values with (*) indicate that they were tested at the 5% significance level

2.3.2 Models

Let be Y a binary response variable classified by Preg (Y = 1) and Open (Y = 0). We assume the Bernoulli distribution with probability π of a success (pregnancy) and the *logit* link function.

For the first model (M1) we include in the linear predictor only the fixed effects (Season, Tech: Technician, SEXED: Type of semen, Type: Type of donor cow, (VH_LC: Status of the recipient cow, FI_D: Fertility index of the donor, FI_R: Fertility index of the recipient, TEEF2: Type of embryo, Embryo_stage: stage of development de embryo, CL_side: Side of corpus luteum on ET day, DFE: Number of days after estrus) and the some interactions (between Season, VH_LC and TEEF2, between TEEF2, Embryo_stage and DFE). Then the linear predictor can be expressed by

$$\begin{aligned} \eta = \log \frac{\pi}{1-\pi} &= \text{Season} * \text{VH\_LC} * \text{TEEF2} + \text{CL\_side} \\ &+ \text{TEEF2} * \text{Embryo\_stage} * \text{DFE} + \text{Type} \\ &+ \text{Sexed} + \text{Tech} + \text{FI\_R} + \text{FI\_D} \end{aligned} \tag{2.7}$$

There is a clear lack of fit of the Bernoulli model with the linear predictor (2.7), model M1, as evidenced by the half normal plot of probability with simulation envelope half-normal plot in Figure 2.6, where many of the deviance residuals are outside of the simulated binomial envelope. A possible explanation may be due to the omitted variables in the linear predictor since the random effects were not taken into account.

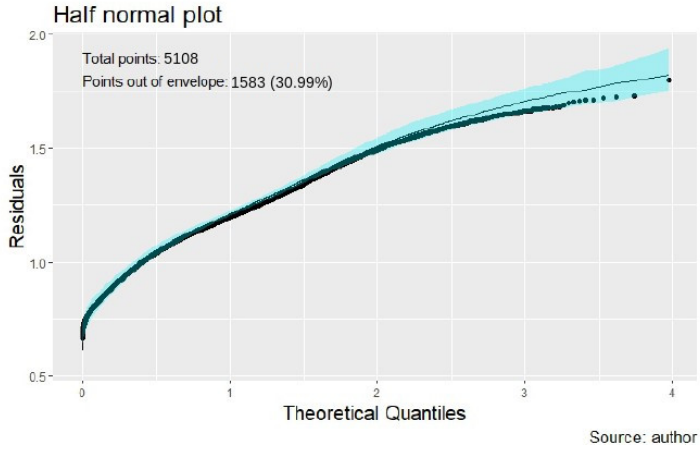

Figure 2.6. Embryo Transfer Data: Half-normal plots of deviance residuals with simulated envelope for model M1

For the second model M2 we add in the linear predictor (2.7) of M1, the random normal effects for donor cow ($N(0, \sigma_d^2)$), recipient cow ($N(0, \sigma_r^2)$) and for sire ($N(0, \sigma_d^2)$), recipient cow ($N(0, \sigma_s^2)$), resulting a Bernoulli-normal model. Other submodels (M3 to M8), using the same fixed effects in the linear predictor as in equation (2.7) but varying the included random effects were fitted and the results are presented in Table 2.1.

From Table 2.1 we see that the -2loglik statistics has the smallest value for models M2 and M3, while the smallest AIC is for M6 and smallest BIC is for M1.

Table 2.1. Results of fitting various models to the ET data, using the same fixed effects in the linear predictor as in Equation 2.7 but varying the included random effects

Model	Fixed	Random	-2loglik	df	AIC	BIC
M1	eq (2.7)	none	6616.823	5041	6750.823	**7188.907**
M2	eq (2.7)	Donor, Recipient, Sire	**6611.747**	5038	6751.747	7209.447
M3	eq (2.7)	Donor, Recipient	**6611.747**	5039	6749.747	7200.908
M4	eq (2.7)	Donor, Sire	6613.140	5039	6751.140	7202.301
M5	eq (2.7)	Recipient, Sire	6614.843	5039	6752.843	7204.004
M6	eq (2.7)	Donor	6613.140	5040	**6749.140**	7193.763
M7	eq (2.7)	Recipient	6614.844	5040	6750.844	7195.466
M8	eq (2.7)	Sire	6616.823	5040	6752.823	7197.445

For comparing the models from Table 2.1 we obtained the differences of the values of -2loglik for the nested models (Table 2.2) and used the likelihood ratio test for testing the components of variance. To test just one component of variance ($H_0 : \sigma_1^2 = 0$), we will compare the calculated value with a mixture of chi-squares (see Equation 3.10) given by

$$\sum_{m=0}^{1} 2^{-1} \binom{1}{m} \chi_m^2 = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2 = 1.921$$

while for two components of variance ($H_0 : \sigma_1^2 = \sigma_2^2 = 0$), is

$$\sum_{m=0}^{2} 2^{-2} \binom{2}{m} \chi_m^2 = \frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2 = 3.419$$

and for three ($H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0$), is

$$\sum_{m=0}^{3} 2^{-3} \binom{1}{m} \chi_m^2 = \frac{1}{8}\chi_0^2 + \frac{3}{8}\chi_1^2 + \frac{3}{8}\chi_2^2 + \frac{1}{8}\chi_3^2 = 4.664.$$

Table 2.2. Difference of values of -2loglik for the nested models from Table 2.1

	M1	M2	M3	M4	M5
M2	5.076 (3)				
M3	5.076 (2)	0.000 (1)			
M4	3.683 (2)	1.393 (1)			
M5	1.980 (2)	3.096 (1)			
M6	3.683 (1)	1.393 (2)	1.393 (1)	0.000 (1)	
M7	1.979 (1)	3.097 (2)	3.097 (1)		0.001 (1)
M8	0.000 (1)	5.076 (2)		3.683 (1)	1.980 (1)

Where (1), (2) and (3) are the number components of variance

The likelihood ratio test show that the best model is M6, the one with linear predictor with Equation 2.7 added by the random effect for donor.

As a next step we tried to reduce the linear predictor for models M3, M6 and M7 testing for the fixed effects. Using the Anova function from the car package, we can see, from Table 2.3, that the effects of Season, Technician (TECH), type of semen (SEXED), status of the recipient

cow (VH_LC), fertility index of donor cow (FI_D), and interactions were not significant. Then the reduced linear predictor is given by

$$\begin{aligned}\eta = \log\frac{\pi}{1-\pi} &= \text{CL\_side} + \text{Type} + \text{FI\_R} + \text{TEEF2} \\ &+ \text{Embryo\_stage} + \text{DFE}. \end{aligned} \tag{2.8}$$

Using the drop1 function from stats package the reduced linear predictor is given by

$$\eta = \log\frac{\pi}{1-\pi} = \text{CL\_side} + \text{Type} + \text{FI\_R}. \tag{2.9}$$

Table 2.3. Analysis of Deviance Table for the Models

		M3	M6	M7
Source	Df	Pr(>Chisq)	Pr(>Chisq)	Pr(>Chisq)
Season	3	0.7093	0.6868	0.7799
Tech	5	0.1329	0.1376	0.1308
Sexed	1	0.5689	0.5577	0.5400
Type	2	0.0403 *	0.0390 *	0.0406 *
VH_LC	1	0.0548	0.0569	0.0556
FI_D	1	0.5715	0.5575	0.6676
FI_R	1	0.0205 *	0.0192 *	0.0199 *
TEEF2	2	0.0000 ***	0.0000 ***	0.0000 ***
Embryo_stage	2	0.0183 *	0.1627	0.0226 *
CL_side	1	0.0307 *	0.0303 *	0.0295 *
DFE	3	0.0000 ***	0.0000 *	0.0000 ***
Season:VH_LC:TEEF2	17	0.3899	0.3741	0.3759
TEEF2:Embryo_stage:DFE	27	0.6759	0.7923	0.7625

P-value significance codes: 0 '***' 0.001 '**' 0.01 '*'

Table 2.4 shows the values of -2loglik statistics, AIC and BIC, considering the reduced linear predictors with Equation 2.8 and Equation 2.9, and different random effects.

Table 2.4. Results of fitting various models to the ET data, using different fixed effects in the linear predictor and varying the included random effects

Model	Fixed	Random	-2loglik	df	AIC	BIC
M9	eq (2.8)	none	6676.744	5096	6700.744	**6779.207**
M10	eq (2.8)	Donor, Recipient	**6672.193**	5094	6700.193	6791.733
M11	eq (2.8)	Donor	6672.995	5095	**6698.995**	6783.996
M12	eq (2.8)	Recipient	6675.499	5095	6701.499	6786.500
M13	eq (2.9)	none	6784.399	5103	6794.399	6827.092
M14	eq (2.9)	Donor, Recipient	6779.558	5101	6793.558	6839.328
M15	eq (2.9)	Donor	6780.369	5102	6792.369	6831.600
M16	eq (2.9)	Recipient	6783.083	5102	6795.083	6834.314

For comparing the models from Table 2.4, we obtained the differences of the values of -2loglik for the nested models (Table 2.5) and used the likelihood ratio test for testing the linear predictors with Equation 2.8 versus Equation 2.9 and also to test the components of variance.

Table 2.5. Difference of values of -2loglik for the nested models from Table 2.4

	M9	M10	M11	M12	M13	M14
M10	4.551 (2)					
M11	3.749 (1)	0.802 (1)				
M12	1.245 (1)	3.306 (1)				
M13	107.655 (7)					
M14		107.365 (7)			4.841 (2)	
M15			107.374 (7)		4.030 (1)	0.811 (1)
M16				107.584 (7)	1.316 (1)	3.525 (1)

Where (1), (2) and (7) are the number components of variance

The likelihood ratio tests show that the best model is M11, the one with linear predictor with Equation 2.8 added by the random effect for donor cow, which also gives the smaller AIC as shown in Table 2.4. The half-normal plot in Figure 2.7 shows evidence of an adequate model with all of the observed deviance residuals lying within the simulated envelope.

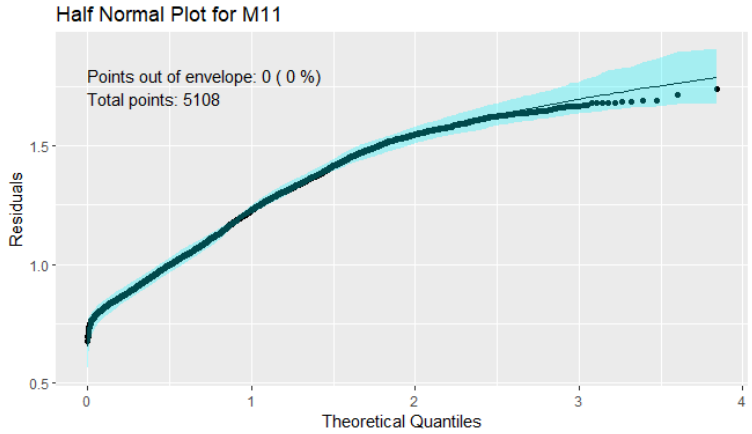

Figure 2.7. Embryo Transfer Data: Half-normal plots of deviance residuals with simulated envelope for model M11

2.3.3 Estimates

The chosen model M11 involves the response variable Y, that can assume the values 0 (Open) or 1 (Preg), that has a Bernoulli distribution with probability π of a success (pregnancy) and the *logit* link function with the linear predictor

$$\begin{aligned} \eta = \log \frac{\pi}{1-\pi} &= \text{CL\_side} + \text{Type} + \text{FI\_R} + \text{TEEF2} \\ &+ \text{Embryo\_stage} + \text{DFE} + Z \end{aligned} \tag{2.10}$$

where Z is the the random effect for donor cow, $Z \sim \text{N}(0, \sigma_d^2)$. The estimates for the parameters for models M9, M10, M11 and M12 are presented in Table 2.6, in order to show the differences in estimation between them. It is important to note that the coefficients with a positive sign contribute to increase the chance of a pregnancy.

Table 2.6. Parameters estimates with their respective standard errors (SE), and components of variance

Param.	M9 Est. (SE)	M10 Est. (SE)	M11 Est. (SE)	M12 Est. (SE)
Intercept	0.2791 (0.0869)**	0.2833 (0.0932)**	0.2775 (0.0917)**	0.2861 (0.0890)**
Type_H	−0.2281 (0.1224)	−0.2370 (0.1293)	−0.2336 (0.1275)	−0.2327 (0.1249)
Type_M	0.0323 (0.0612)	0.0343 (0.0665)	0.0327 (0.0657)	−0.0344 (0.0624)
FI_R	0.0449 (0.0197)*	0.0459 (0.0202)*	0.0452 (0.0198)*	0.0456 (0.0201)*
CL_side_R	0.1324 (0.0589)*	0.1348 (0.0604)*	0.1326 (0.0594)*	0.1350 (0.0602)*
TEEF2_1	0.4292 (0.0759)***	0.4446 (0.0803)***	0.4362 (0.0786)***	0.4400 (0.0782)***
TEEF2_2	−0.0343 (0.0851)	−0.0373 (0.0885)	−0.0377 (0.0872)	−0.0338 (0.0869)
Emb_st_5	0.1705 (0.0714)*	0.1824 (0.0740)*	0.1815 (0.0729)*	0.1719 (0.0728)*
Emb_st_6	0.3069 (0.1434)*	0.3255 (0.1480)*	0.3204 (0.1459)*	0.3135 (0.1461)*
DFE_6	−0.3052 (0.0756)***	−0.3008 (0.0779)***	−0.2976 (0.0766)***	−0.3089 (0.0772)***
DFE_8	−0.0419 (0.0699)	−0.0397 (0.0721)	−0.0398 (0.0711)	−0.0416 (0.0713)
DFE_9	−0.5727 (0.1425)***	−0.5891 (0.1471)***	−0.5822 (0.1446)***	−0.5813 (0.1456)***
AIC	6700.744	6700.193	**6698.995**	6701.499
-2loglik	6676.744	**6672.193**	6672.995	6675.499
σ^2_d	–	0.0369	0.0383	–
σ^2_r	–	0.0695	–	0.0863

Significance ($^{***}p-value<0.001, ^{**}p-value<0.01, ^{*}p-value<0.05$);
σ^2_d (donor_cow) and σ^2_r (recipient)

The recipient fertility index was significant in model M11 from Table 2.6, confirming what was seen in Figure 2.1 and the positive coefficient shows that as fertility index increases, the chance of a positive result of pregnancy after ET increases.

Table 2.6 also shows evidence of significant differences in pregnancy rates for CL_side, and embryo stage. Fresh grade 1 embryos have a higher chance of pregnancy success when compared to grade 2 fresh or frozen. Embryos transferred 7 and 8 days after estrus have a higher change of pregnancy success than those transferred 6 and 9 days.

2.4 Final remarks

Embryo transfer is a powerful technology for bovine genetic improvement, primarily to propagate the genes of females with superior genetic values and lineage. Many variables can affect pregnancy outcomes. High pregnancy rates can be achieved with good quality embryos, transferred by an experienced embryo transfer technician, to well selected and managed recipients 7 or 8 days after estrus. Embryo transfer studies are usually not planned experiments but observational data which are very unbalanced and heterogeneous.

The aim of this work was to analyze data from ET, in which the Preg (Y = 1) and Open (Y = 0) outcomes described in subsection 2.2.1. Statistical analysis for binary data can be performed using the Bernoulli distribution, which has the assumption of constant probability of success, and is a particular case of a GLM (Nelder e Wedderburn, 1972).

The convenience of using GLMs does not necessarily imply goodness of fit. The model M1 gave a lack of fit to the ET data, probably due to variables not included in the model, causing erroneous interpretation. Moreover, the probability of success was not constant in this study, since the occurrence of pregnancy is influenced by characteristics related both to the embryo and the recipient cow, that is, some females are more likely to be pregnant than others.

Several models were fitted to the ET incorporating one or more random effects at the levels of donor cow, recipient cow and sire in the linear predictor, to try to accommodate possible sources of variation and correlation, with a Bernoulli-logistic-normal model being proposed. The absence of significance in some covariates can be explained by the presence of multicollinearity between them (Gujarati e Porter, 2011).

There is not a recipe for fitting statistical models for this type of dataset. Every dataset presents its own peculiarities which may guide the statistical analyses one way or another.

ACKNOWLEDGMENT

The first author acknowledges the scholarship financial support from CNPq, Brazil. This research was partially supported by CNPq, Brazil, for CGBD.

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademinai Kiado.

Bates, D., Mächler, M., Bolker, B., e Walker, S. (2015). Fitting linear mixed-effects models using lme4. R package version 3.3.1.

de Andrade Moral, R., Hinde, J., e Garcia Borges Demétrio, C. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(10).

Gujarati, D. N. e Porter, D. C. (2011). *Econometria Básica*. Amgh Editora, 5 edition.

H.A (2017). Updates to the Total Performance Index (TPI) and Type Composites. *Holstein Association USA*.

Hansen, P. J. (2020). The incompletely fulfilled promise of embryo transfer in cattle—why aren't pregnancy rates greater and what can we do about it? *Journal of Animal Science*, 98(11):skaa288.

Hinde, J. e Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

Jørgensen, B. B. (2002). Generalized linear models.

McCullagh, P. e Nelder, J. A. (1989). *Generalized linear models*, volume 37. Chapman and Hall/CRC, New York, 2 edition.

McCulloch, C. E. e Searle, S. R. (2001). *Generalized, linear, and mixed models*, volume 1. John Wiley & Sons, New York.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2017). Hierarchical models with normal and conjugate random effects: a review. *SORT-Statistics and Operations Research Transactions*, 1(2):191–254.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, 25(3):325–347.

Nelder, J. e Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135:370–384.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Self, S. G. e Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Reading, MA.

Verbeke e Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Viana, J. H. M. (2022). 2022 statistics of embryo production and transfer in domestic farm animals. *IETS Data Retrieval Committee. Embryo Technology Newsletter*, 40(4).

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59.

Wright, J. (2001). Appendix 1: Photographic illustrations of embryo developmental stage and quality codes. *Manual of the International Embryo Technology Society*, pages 1–4.

3 A NEW APPROACH TO EMBRYO VIABILITY DATA ANALYSIS USING COMBINED MODELS

Abstract

Generalized Linear Models are extensively used to analyze proportion data. However, correlation structures and the presence of overdispersion in the data cannot be accommodated by the standard binomial model. Thus, extensions such as the Binomial-logistic-normal hierarchical model and the combined Binomial-logistic-normal-beta model are used in this work to take into account overdispersion and correlation. The motivation for this work was an embryo production data for which the response variable of interest was the ratio of the number of viable embryos by the of total embryos, defined as the embryo viability rate. The aim was to identify the factors that influence the production of the viable embryo in vivo. The main results suggest that the combined model gave the best fit, and was able to capture the extra variability and correlation, satisfactorily.

Keywords: Binomial Distribution; Overdispersion; Viable Embryo.

3.1 Introduction

Superovulatory treatment is the crucial phase to the viability of the embryo production in vivo technique, and embryo transfer (ET) (Bó e Mapletoft, 2013). This is characterized by the intramuscular or subcutaneous application of decreasing doses of follicle stimulating hormone (FSH) in the animal, being administered twice daily throughout the process (Laster, 1972). Naturally, in a fertile bovine female, every 21 days, only one oocyte is released to fertilization, that is, featuring low annual reproductive rate (Sartori et al., 2009). Thus, the main objective of the superovulation technique is to increase the oocytes number. Consequently, the amount of embryos that can be collected and transferred is also increased.

Although considerable progress has been made in reproductive efficiency of bovine females, and subsequently improved genetic utilization of the herd, most of the embryos produced, later, will not be viable for the embryo transfer process. The probability of success for a viable embryo is influenced by many factors related to the donor cow (reproductive history, fertility index, somatic cell count and milk production) and the type of semen of the sire donor. Moreover, environmental factors also directly affect these reproductive traits, such as the seasons of the year (Thatcher et al., 2001; Sartori et al., 2002a; Hansen, 2009; Wiltbank et al., 2006; Andreu-Vázquez et al., 2012).

Given the number of factors involved, the unpredictability of the viable embryo numbers is one of the main obstacles for the embryo transfer technique. However, it is still unknown how each of these factors contributes to the variability observed in the response to the superovulation, and in the production of embryos. Therefore, it is important the identification of which of these effects are really contributing to provide information and make possible the increase in the efficiency of the embryo production in vivo technique, and consequently in embryonic viability.

The binomial distribution, a member of the exponential family and a particular case of

a generalized linear model (Nelder e Wedderburn, 1972), is a starting point for the analysis of proportion data. For this very simple model we assume that events happen independently, singly, and at random with some constant probability of success. However, the embryo proportions are overdispersed, and the variability of the data is larger than the variability specified by the binomial model (Hinde e Demétrio, 1998). The response variable number of viable embryos out of a total number of embryos have two extra-variability sources, one caused by the heterogeneity of the animals (overdispersion) and the other by the hierarchical structure caused by random effects of donor and sire (correlation between observations) (Lindsay, 1986; Breslow e Clayton, 1993; Molenberghs et al., 2012).

Many different specific models for overdispersion can arise from alternative possible mechanisms for the underlying process (Hinde e Demétrio, 1998). The simplest way is to assume some more general form for the variance function, possibly including additional parameters, leading to the quasi-binomial model. Another way is to assume a two-stage model for the response, that is, to assume that the basic response model parameter itself has some distribution having as a typical example the beta-binomial model. An alternative model arises from the inclusion of random effects in the linear predictor of the model as the Binomial-normal model an example of a generalized linear mixed model (GLMM), allowing to get a measure of intraclass correlation.

Furthermore, overdispersion and correlation between observations may occur simultaneously, and models accommodating them at once are less than common. A combined model can be used to accommodate overdispersion and clustering through two separate sets of random effects, of gamma and normal type, respectively (Molenberghs et al., 2007, 2010).

The aim of this paper is to propose models that best fit to the embryonic viability rate with a binomial response, and describe ways to analyze these type of data with the goal of identifying which factors influence the viability of embryos.

In this work we review and compare methods for analyzing proportion data with particular focus on potential applications in agricultural research. In subsection 3.2.1 we provide a motivation data set. In subsection 3.2.3 we present some models used for the analysis of proportion data, while subsection 3.2.4 discusses model selection and diagnostics. The motivation data set is analyzed in section 2.3. Some general considerations are presented in section 2.4. The scripts developed in the software **R** (R Core Team, 2018) are presented in the Appendix.

3.2 Material and Methods

3.2.1 Case Study: Embryo Production Data

To better understand the genetic indexes effect, obtained by genomic evaluation, and other possible factors that interfere on the viable embryo production rate, observational studies were performed at Ruann and Maddox Dairy Farm in Riverdale, California (USA), between years from 2012 to 2018. The response variable consisted of the ratio of the number of viable embryos by the total number of embryos, per donor cow, characterizing the embryo viability rate.

For the selection of animals were considered the ones with the highest genetic values of

the breed, using 454 donor cows and 261 donor sires, repeating the best crosses (donor cow × donor sire), totaling 1457 observations.

The embryo production was performed in vivo, a process in which the donor females were submitted to the superovulation protocol based on the use of the follicle stimulating hormone (FSH), started between days 7 and 10 after estrus detection. Artificial insemination (AI) was performed between 12 and 24 hours after the first detection of estrus, according to the guidelines of the Manual of the International Embryo Transfer Society (IETS, 2010).

All the collected embryos were obtained non-surgically by Sugie's adapted system (Kanagawa et al., 1995). This closed system consists of equipment with three way that attaches one liter of DPBS to perform uterine lavage, a millipore® filter to retain the embryos, and a two-way catheter, which will be fixed in the uterus to wash. After epidural anesthesia of the donor, the catheter is fixed in the final third of the right uterine horn, where approximately 10 washes are performed with 50 ml of DPBS. Repeat the same procedure on the other horn.

The specialized filter for the transport of the embryos was sent after collection to the laboratory. The structures found were put in a maintenance solution and maintained at environments temperatures. The number of embryos recovered were counted, and those considered viable, separated and evaluated, for later use in the embryo transfer process. Only morulae and blastocysts classified with grade I, II and III were considered viable (Wright, 1981; Thompson et al., 1998).

Additionally, it was obtained the information of fertility index of donor cow (FI); total donor cow performance index (TPI); daughters pregnancy rate of donor cow (DPR); donor cow somatic cell score (SCC: $\times 1,000/ml$); milk production (Milk - milk averages in pounds); life stage of the donor cow (TYPE: 1 - dry cow (D: 587), 2 - heifer (H: 198) and 3 - lactanting cow (M: 672)); category of cow response to stimulation of superovulation (TYPE_2: 1 - Good (616), 2 - Big (64), 3 - Fair (550) and 4 - Poor (227)); embryo viability rate (viable embryos number / total embryos produced number); type of semen (SEXED: 1 - not sexed (1242) and 2 - sexed (215)); donor cow and donor sire.

In Fig. 3.1, we have the scatter plot of the proportions of viable embryos versus each of the continuous explanatory variables (FI, TPI, DPR, Milk, SCC) of the donor cow.

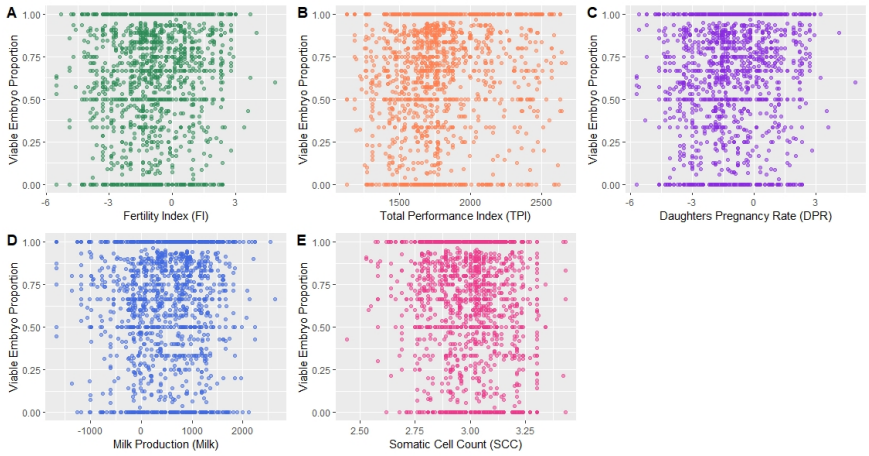

Figure 3.1. Embryo Production Data: descriptive graph for the variables referring to the donor cow that are continuous in nature

The data of the embryo production were recorded and divided by seasons of the year: 1 - Summer (279: July, August, September); 2 - Autumn (374: October, November, December); 3 - Winter (385: January, February, March) and 4 - Spring (419: April, May, June).

3.2.2 Exploratory analysis

In order to understand the generating process of the data and and to visualize possible interactions between factors, before fitting models some descriptive analysis are presented. For the descriptive analysis of the data, an initial examination of the dispersion plots of the embryo viability rate by i) fertility index considering somatic cell score, ii) response category, per type of semen used, iii) total number of embryos produced versus viable embryos number, iv) average of the fertility index versus the average embryo viability rate, per season of the year and the donor cow's life stage. In addition, the quantitative variables of the data set, such as the age cow, somatic cell score (SCC), total performance index (TPI), daughter pregnancy rate (DPR), fertility index (FI) and milk production (Milk), were used to obtain the coefficients and graphs of simple and partial correlations (Demétrio e Zocchi, 2011).

For descriptive evaluation, the graphs are presented in Figure 3.2. It is observed that in the scatter plot (Figure 3.2A), there is no evidence of a relationship between the proportion of viable embryos and the total number of embryos, that is, the proportion of embryonic viability does not depend on the number of embryos obtained at fertilization. In addition, in the Figure 3.2B, the variability found indicated that there were differences in the embryo viability rate, regarding the response categories. And when we using non sexed semen, apparently, there were better results in the response variable. At Figure 3.2C, we have different behaviors observed throughout the seasons in relation to the averages of the fertility indexes and the life stages of the donor cow.

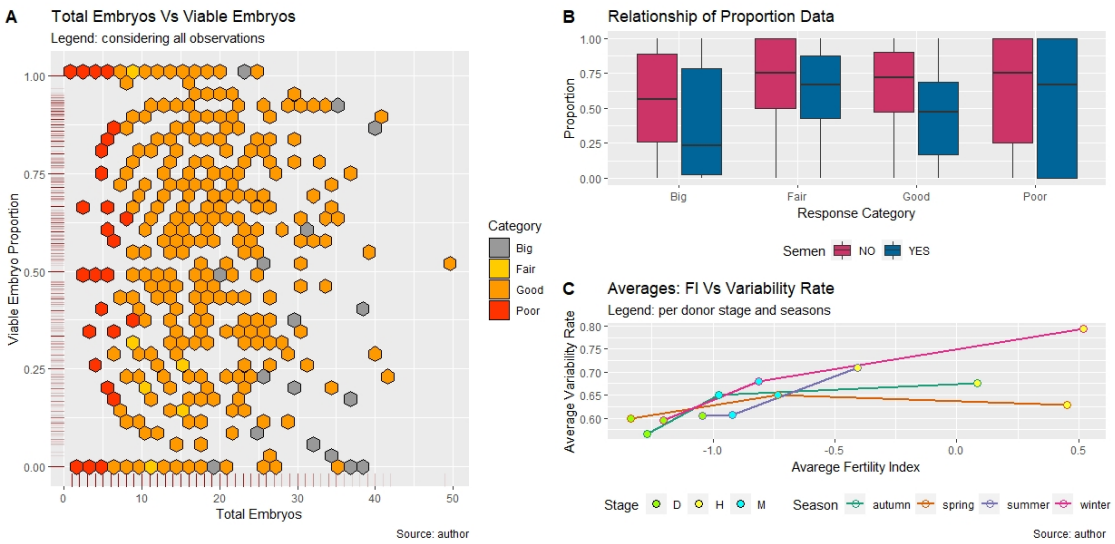

Figure 3.2. Embryo Production Data: A) Dispersion plot between the variables total number of produced embryos and proportion of viable embryos; B) Box-plot of the response category and embryo viability rate, by type of semen used; C) Averages of fertility indexes and embryo viability rate, by seasons and life stages of the donor cow

The estimates of simple and partial correlations can be visualized in the Figure 3.3. According to the coefficients there is a high, positive and significant relationship between the FI and DPR variables, which can be explained by the inclusion of DPR in the calculation of the fertility index. The TPI would be the percentage of milk production and characteristics of the cow. Thus, the correlation between this covariate and that related to milk production (Milk) is expected (H.A, 2017). On the other hand, the simple correlation between the FI and SCC variables showed that there is a negative relationship between them. However, when we removing possible influences from external factors, it can be noted that there is a positive association. In Figure 3.4, there is the dispersion plot between FI and SCC wherein the presence of external factors, apparently, as the somatic cell score decreases, the fertility indexes of the donor cow increase. Moreover, the densities related to the respective variables are found on the margins of the graph.

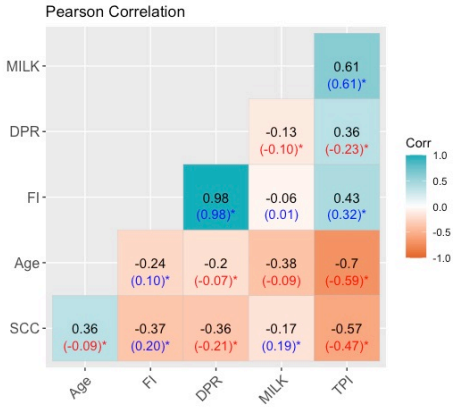

Figure 3.3. Embryo Production Data: Simple correlations (values in black) and partial (values in colors: red - negative associations; blue - positive associations)

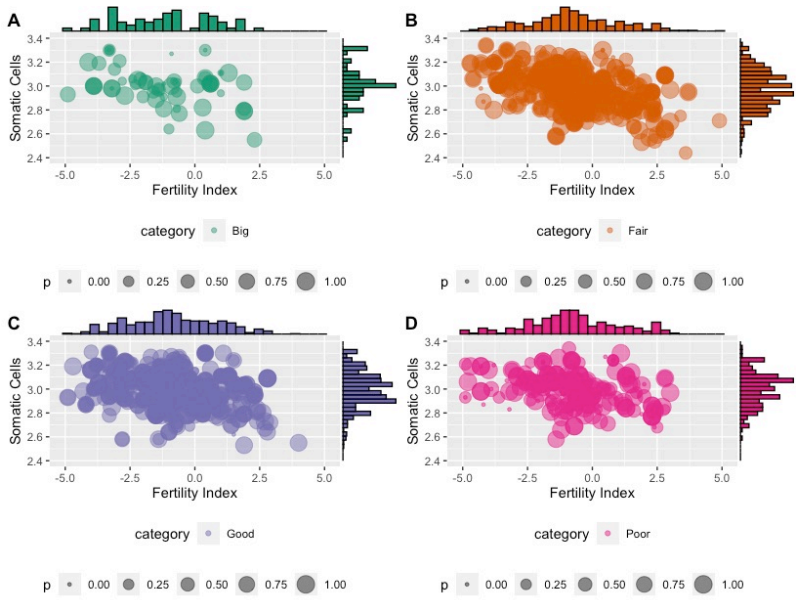

Figure 3.4. Embryo Production Data: Dispersion plot between fertility index (FI) versus somatic cell score (SCC) considering the response category

3.2.3 Statistical Approach: Combined Models Structure

Molenberghs et al. (2010, 2017) propose an extension of the generalized linear model, called combined model, to accommodate overdispersion and clustering through two separate sets of random effects, of gamma and normal type, respectively. This considers the hierarchical data structure (observations that are nested by some effect).

Let Y_{ij} be a random variable with *j-th* clustered measurement ($j = 1, \cdots, n_i$) of the *i-th* element ($i = 1, \cdots, N$), and supposing conditionality over random effects $\boldsymbol{u} \sim \mathrm{N}_{\boldsymbol{q}}(\mathbf{0}, \boldsymbol{\Sigma})$, we have the expression

$$f_i(y_{ij}|u, \beta, \theta_{ij}, \phi) = \exp\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\}, \tag{3.1}$$

with conditional mean given by

$$E(Y_{ij}|u_i, \beta, \theta_{ij}) = \mu_{ij}^c = \psi(\lambda_{ij}) = \theta_{ij}\kappa_{ij}, \tag{3.2}$$

where $\theta_{ij} \sim \Theta_{ij}(\upsilon_{ij}, \sigma_{ij}^2)$ with mean υ_{ij} and variance σ_{ij}^2 and $\kappa_{ij} = g(x_{ij}^T\beta, z_{ij}^T u)$ (Molenberghs et al., 2010).

Assuming that the data comes from binary measures, Bernoulli distribution with logistic link function is considered for the response variable. Furthermore, for random effects, the normal distribution is assumed to accommodate the correlation structure in the longitudinally measured observations, and the beta distribution to accommodate overdispersion. Thus, we have the Bernoulli-logistic-normal-beta model given by

$$\begin{aligned} Y_{ij}|\theta_{ij}, u_i &\sim \text{Bernoulli}(\pi_{ij}) \\ \pi_{ij} &= \theta_{ij}\frac{\exp(x_{ij}^T\beta + z_{ij}^T u)}{1 + \exp(x_{ij}^T\beta + z_{ij}^T u)} \\ \theta_{ij} &\sim \text{Beta}(a, b) \\ u_i &\sim \mathrm{N}_q(0, \Sigma), \end{aligned} \tag{3.3}$$

where Σ is the variance and covariance matrix of the q dimensional vector of u_i.

However, when we have count data it is appropriate to assume the Poisson distribution for response variable with logarithmic link function. In addition, for random effects, the normal distribution is assumed to accommodate correlation and gamma distribution to capture overdispersion. Thus, the Poisson-normal-gamma model is defined as

$$\begin{aligned} Y_{ij}|\theta_{ij}, u_i &\sim \text{Poisson}(\pi_{ij}) \\ \pi_{ij} &= \theta_{ij}\exp(x_{ij}^T\beta + z_{ij}^T u) \\ \theta_{ij} &\sim \Gamma(a, b) \\ u_i &\sim \mathrm{N}_q(0, \Sigma), \end{aligned} \tag{3.4}$$

where a and b are the gamma distribution parameters.

The estimation of the parameters is obtained by maximizing the logarithm of the likelihood function. Thus, considering binary data, and assuming the distribution is Bernoulli, we have the probability mass function of Y_{ij} conditional on θ_{ij}, u_i expressed by

$$\begin{aligned} f(y_{ij}|\theta_{ij}, u_i) &= \left[\theta_{ij}\frac{\exp(x_{ij}\beta + z_{ij}u)}{1 + \exp(x_{ij}\beta + z_{ij}u)}\right]^{y_{ij}} \\ &\times \left[1 - \theta_{ij}\frac{\exp(x_{ij}\beta + z_{ij}u)}{1 + \exp(x_{ij}\beta + z_{ij}u)}\right]^{1-y_{ij}} \end{aligned} \tag{3.5}$$

where β is vector of parameters of fixed effects. Considering the multivariate Gaussian distribution for the random effects u_i in the linear predictor , and the beta distribution for the random effect θ_{ij}, we have the joint density function for the contribution of the *i-th* individual given by

$$f(y_i, \theta_i, u_i) = \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, u_i) f(u_i) f(\theta_{ij}), \tag{3.6}$$

where conditional independence is assumed, that is, since the dependencies on the random effects u_i and θ_{ij} were included in the likelihood function.

In this context, the estimation of $\beta's$ and the components of variance of matrix Σ by maximum likelihood, will be obtained integrating the (Equation 3.6) in the parametric space of u_i and θ_{ij}, in relation to the expression given by

$$f(y_i, \theta_i) = \int \prod_{j=1}^{n_i} f(y_{ij}|\theta_{ij}, u_i) f(u_i) f(\theta_{ij}) du_i. \tag{3.7}$$

According Molenberghs et al. (2010) the likelihood function for the Bernoulli-logistic-normal-beta model, conditioned to the random effect vector θ and with N individuals, will be

$$\begin{aligned} L(\beta, \Sigma, a, b|\theta) &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \left[\theta_{ij} \frac{\exp(x_{ij}\beta + z_{ij}u_i)}{1 + \exp(x_{ij}\beta + z_{ij}u_i)}\right]^{y_{ij}} \\ &\times \left[1 - \theta_{ij} \frac{\exp(x_{ij}\beta + z_{ij}u_i)}{1 + \exp(x_{ij}\beta + z_{ij}u_i)}\right]^{1-y_{ij}} \\ &\times \frac{1}{\sqrt{(2\pi)^{n_i}}} \frac{1}{\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} u_i^T \Sigma^{-1} u_i\right) \\ &\times \frac{\theta_{ij}^{a-1}(1-\theta_{ij})^{b-1}}{B(a,b)} du_i. \end{aligned} \tag{3.8}$$

Finally, after considering $y_{ij} = 0$ and $y_{ij} = 1$, in cases where there are binary data (Rizzato, 2011), we obtain the marginal likelihood function conditioned only to the random effect of the linear predictor u_i as

$$\begin{aligned} L(\beta, \Sigma, a, b) &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \left(\frac{ak_{ij}}{a+b}\right)^{y_{ij}} \left[\frac{(1-k_{ij})a + b}{a+b}\right]^{1-y_{ij}} \\ &\times f(u_i|\Sigma) du_i \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \frac{1}{a+b} (ak_{ij})^{y_{ij}} [(1-k_{ij})a + b]^{1-y_{ij}} \\ &\times f(u_i|\Sigma) du_i. \end{aligned} \tag{3.9}$$

The Equation 3.9 does not allow to obtain the estimators analytically, being carried out by means numerical methods. A possible solution to the integration will be the Gauss-Hermite adaptive quadrature algorithm. Once the convergence criteria is achieved, the estimates obtained can be used for statistical inference.

Model selection - Inference for components of variance

To test random terms, useful asymptotic test is the likelihood-ratio test (LR) that are based on comparing the values of likelihood functions of two nested models, having the same

set of fixed-effect parameters, but different sets of covariance parameters. The likelihood ratio statistics is given by

$$LR = -2[\text{logLik(reduced model) - logLik(complete model)}],$$

where logLik is the logarithm of the likelihood function.

When there is no parameter on the boundary of the parametric space, $LR \sim \chi^2_\nu$, where ν is the difference in number of degrees of freedom between the two models. LR has a distribution that is a mixture of χ^2's when there are parameters on the boundary of the parametric space (Self e Liang, 1987). In the case of component of variance models with independence between the random effects as in our case the mixture of χ^2's is given by

$$\sum_{m=0}^{k'} 2^{-k'} \binom{k'}{m} \chi^2_m. \tag{3.10}$$

Akaike (AIC) and Bayesian (BIC) Information Criteria are used when the two models being compared are non nested AIC (AKAIKE, 1973; Schwarz, 1978),

$$\text{AIC} = -2\text{logLik} + 2p,$$

and

$$\text{BIC} = -2\text{logLik} + \log(n)p.$$

where p is the number of fitted parameters and n is the number of observations. The best model is considered the one that presents the lowest values of AIC and/or BIC.

Model selection - Inference for fixed effects

After choosing a model for random terms, to test for fixed effects we can use the likelihood-ratio test (LR) that are based on comparing the values of likelihood functions of two nested models, having the same set of random-effect parameters, but different sets of fixed parameters

Akaike (AIC) and Bayesian (BIC) Information Criteria can be used when the two models being compared are non nested (AKAIKE, 1973; Schwarz, 1978).

To help with the selection of a model the drop1 function from stats package computes all the single terms in the scope argument that can be added to or dropped from the model, fit those models and compute a table of the changes in fit. It gives a comparison of models based on the AIC criterion and when using the option test="F" adds a "type II ANOVA" (using Anova function from the car package). The hierarchy is respected when considering terms to be added or dropped: all main effects contained in a second-order interaction must remain, and so on.

3.2.4 Statistical Approach: Data Modeling

Models

Let Y be a random variable given by the number of viable embryos out of the total (m) number of embryos. The standard distribution assumed is the binomial with probability π of a success (occurrence of a viable embryo) and the *logit* link function.

A reason for extending the binomial model is because of the occurrence of a hierarchical structure in the data caused by a clustering resulted from repeatedly measuring the outcome on the same experimental unit (e.g. embryos obtained with semen from the same sire or oocytes from the same cow). The possible correlation between measurements for the same individual is often accommodated through the inclusion of subject-specific, random effects (Verbeke e Molenberghs, 2000). We considered the inclusion of a random effect in the linear predictor with logistic link function, leading to the binomial-logistic-normal (Williams, 1982), expressed by

$$\begin{aligned} Y_{ijklrs}|u_i &\sim \text{Binomial}(m_{ijklrs}, P_{ijklrs}) \\ P_{ijklrs} &= \frac{\exp(x^T_{ijklrs}\beta + z^T_{ijklrs}u)}{1+\exp(x^T_{ijklrs}\beta + z^T_{ijklrs}u)} \\ u_i &\sim \text{N}(0, \Sigma), \end{aligned}$$

and linear predictor

$$\begin{aligned} \eta_{ijklrs} = \log\left(\frac{P_{ijklrs}}{1-P_{ijklrs}}\right) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \\ &+ \beta_3 X_{3i} + \beta_4 X_{4i} + \\ &+ \beta_5 X_{5i} + \beta_6 X_{6i} + \tau_j + \\ &+ \delta_k + \varphi_l + \alpha_r + \gamma_s + \\ &+ \xi_{i(jklrs)}, \qquad \text{(M1)} \end{aligned}$$

where β_0 is the fixed effect constant, β_h with $(h = 1, \cdots, 6)$ are, respectively, the regression coefficients associated with the quantitative variables: X_{1i} fertility index of *i-th* fixed effect donor cow, X_{2i} is the somatic cell score of *i-th* fixed effect donor cow, X_{3i} is the total performance index of *i-th* fixed effect donor cow, X_{4i} is the pregnancy rate of the daughters of *i-th* fixed effect donor cow X_{5i} is the milk production in pounds of *i-th* fixed effect donor cow, X_{6i} is the age of *i-th* fixed effect donor cow. Moreover, τ_j is the fixed effect of *J-th* year $(j = 1, 2, 3, 4, 5, 6, 7, 8)$, δ_k is the fixed effect of *k-th* season $(k = 1, 2, 3, 4)$, φ_l is the fixed effect of *l-th* life stage of donor cow $(l = 1, 2, 3)$, α_r is the fixed effect of *r-th* response category $(r = 1, 2)$, γ_s is the fixed effect of *s-th* type of semen $(s = 1, 2)$, and $\xi_{i(jklrs)}$ is the random effect at crossover level, with $\xi_{i(jklrs)} \sim \text{N}(0, \sigma^2_c)$.

Overdispersion and correlation between observations may occur simultaneously, and models accommodating both at once were proposed through two separate sets of random effects, of gamma and normal type, respectively (Molenberghs et al., 2007, 2010, 2017) leading to an unified modeling framework, termed the combined model. Thus, the second model proposed was the Binomial-logistic-normal-beta distribution, described by

$$\begin{aligned} Y_{ijklrs}|\theta_{ijklrs}, u_i &\sim \text{Binomial}(m_{ijklrs}, P_{ijklrs}) \\ P_{ijklrs} &= \theta_{ijklrs}\frac{\exp(x^T_{ijklrs}\beta + z^T_{ijklrs}u)}{1+\exp(x^T_{ijklrs}\beta + z^T_{ijklrs}u)} \\ \theta_{ijklrs} &\sim \text{Beta}(a, b) \\ u_i &\sim \text{N}_q(0, \Sigma), \end{aligned}$$

and linear predictor

$$\begin{aligned}\eta_{ijklrs} = \log\left(\frac{P_{ijklrs}}{1-P_{ijklrs}}\right) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \\ &+ \beta_3 X_{3i} + \beta_4 X_{4i} + \\ &+ \beta_5 X_{5i} + \beta_6 X_{6i} + \tau_j + \\ &+ \delta_k + \varphi_l + \alpha_r + \gamma_s + \\ &+ \xi_{i(jklrs)}, \qquad \text{(M2)}\end{aligned}$$

where the parameters of fixed and random effects were defined in model M1.

Based on the structure of the model M2, we have the following submodels

$$\begin{aligned}\eta_{iklrs} = \log\left(\frac{P_{iklrs}}{1-P_{iklrs}}\right) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \\ &+ \beta_3 X_{3i} + \beta_5 X_{5i} + \delta_k + \\ &+ \varphi_l + \alpha_r + \gamma_s + \\ &+ \xi_{i(klrs)}, \qquad \text{(M3)}\end{aligned}$$

$$\begin{aligned}\eta_{ikrs} = \log\left(\frac{P_{ikrs}}{1-P_{ikrs}}\right) &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \\ &+ \beta_3 X_{3i} + \delta_k + \varphi_l + \alpha_r + \\ &+ \gamma_s + \xi_{i(krs)}, \qquad \text{(M4)}\end{aligned}$$

where the parameters were defined in the model M1.

Estimation of the parameters

All the analyses were implemented in the software R (R Core Team, 2018). The parameters estimation of the hierarchical model M1 was performed by means of the maximum likelihood, obtained by the iterative method of Gauss-Hermite adaptive quadrature, which is implemented in the function glmer of package lme4 (Bates et al., 2015). On the other hand, the parameter estimates of the combined models M2, M3 and M4, were obtained by the restricted maximum likelihood method, which is implemented in the glmmTMB (Brooks et al., 2017). All significance of the estimated effects were based at the 5% level.

For the comparisons between the models, we used the deviance statistic (-2ℓ), where ℓ is based on the maximized logarithmic value of the likelihood. In addition, we used the Generalized Akaike Information Criteria (AIC) (Cordeiro e Demétrio, 2008).

Goodness of fit and diagnostics

After fitting a model to a data set, in addition to the global goodness-of-fit, it is useful to use some diagnostic plots to detect specific aspects of possible model failure.

To check that the residuals are consistent with the variation implied by the model, an approach is to use the half normal plot with simulated envelope which is implemented in the hnp package (de Andrade Moral et al., 2017).

3.3 Results

In the Table 3.1, we have models proposed in accordance with those described in subsection 3.2.4. For the M1 model, the AIC value showed that the adjustment was lower when compared to the others. When adjusting the M2 model, distributions combinations, it is noted that both the AIC and the deviance statistics were smaller. While the quality of the fit was considerably better when compared to M1, as shown by Figure 3.5A and Figure 3.5B. Thus, based on the adjustment M2 model covariates that did not present evidence of significance were excluded, resulting in the M3 and M4 models. In the comparison of the reduced models, it highlighted the similarity between the values of the selection criteria (AIC and deviance). However, the verification of the goodness of fit shows that the simplest model will be the most indicated (Figure 3.5).

Table 3.1. Parameters estimates with their respective standard errors (SE), and variance components to proposed models

Parameters	M1 Estimate (SE)	M2 Estimate (SE)	M3 Estimate (SE)	M4 Estimate (SE)
β_0	1.6405 (1.3940)	0.9436 (0.7480)	2.8870 (0.8283)***	2.7200 (0.7558)***
β_1	0.1910 (0.1639)	0.1547 (0.1150)	0.2446 (0.1161)*	0.2855 (0.1111)*
β_2	−0.5604 (0.3856)	−0.4152 (0.2577)	−0.7231 (0.2624)**	−0.6891 (0.2552)**
β_3	−0.1584 (0.1581)	−0.1326 (0.1121)	−0.1927 (0.1129)	−0.2412 (0.1069)*
β_4	0.0008 (0.0003)**	0.0006 (0.0005)	–	–
β_5	−0.0002 (0.0001)	−0.0006 (0.0006)	−0.0001 (0.0001)	–
β_6	−0.0440 (0.0412)	−0.0344 (0.0277)	–	–
τ_2	−0.7021 (0.3552)*	−0.2536 (0.2806)	−0.2212 (0.2823)	–
τ_3	−0.2783 (0.3114)	0.1108 (0.2633)	0.1793 (0.2649)	–
τ_4	−0.4644 (0.3132)	−0.1020 (0.2581)	−0.0467 (0.2587)	–
τ_5	−0.6351 (0.3126)*	−0.2495 (0.2528)	−0.2035 (0.2519)	–
τ_6	−0.7161 (0.3176)*	−0.2370 (0.2543)	−0.1700 (0.2537)	–
τ_7	−0.6378 (0.3357)	−0.2621 (0.2636)	−0.1786 (0.2623)	–
δ_1	−0.3000 (0.1302)*	−0.2212 (0.1060)*	−0.2264 (0.1064)*	−0.2147 (0.1059)*
δ_2	−0.1926 (0.1174)	−0.2388 (0.0966)*	−0.2399 (0.0968)*	−0.2141 (0.0956)*
δ_4	−0.2522 (0.1144)*	−0.1386 (0.0935)	−0.1435 (0.0968)	−0.1467 (0.0936)
φ_1	−0.0523 (0.2704)	0.0127 (0.1966)	−0.3533 (0.1403)*	−0.4460 (0.1314)*
φ_3	−0.0243 (0.2249)	0.0035 (0.1642)	−0.2688 (0.1317)*	−0.3150 (0.1279)*
α_1	0.6690 (0.1925)***	0.3871 (0.1559)*	0.4025 (0.1567)*	0.4079 (0.1563)**
α_3	0.8512 (0.1991)***	0.5490 (0.1591)***	0.5697 (0.1601)***	0.5716 (0.1602)***
α_4	0.4639 (0.2364)*	0.3320 (0.1851)	0.3575 (0.1862)	0.3750 (0.1862)*
γ_2	−1.2789 (0.1367)***	−0.8856 (0.1069)***	−0.8626 (0.1075)***	−0.8953 (0.1040)***
AIC	6183.0690	6102.9000	6036.1000	6034.3000
LogLik	−3068.5345	−3004.3000	−3002.1000	−2996.1000
DEV	6137.0690	6008.6000	6004.2000	5992.2000
σ_c^2	2.1778	0.0112	0.0174	0.0214

Significance ($^{***}p-value<0.001$, $^{**}p-value<0.01$, $^{*}p-value<0.05$) and DEV (deviance of model)

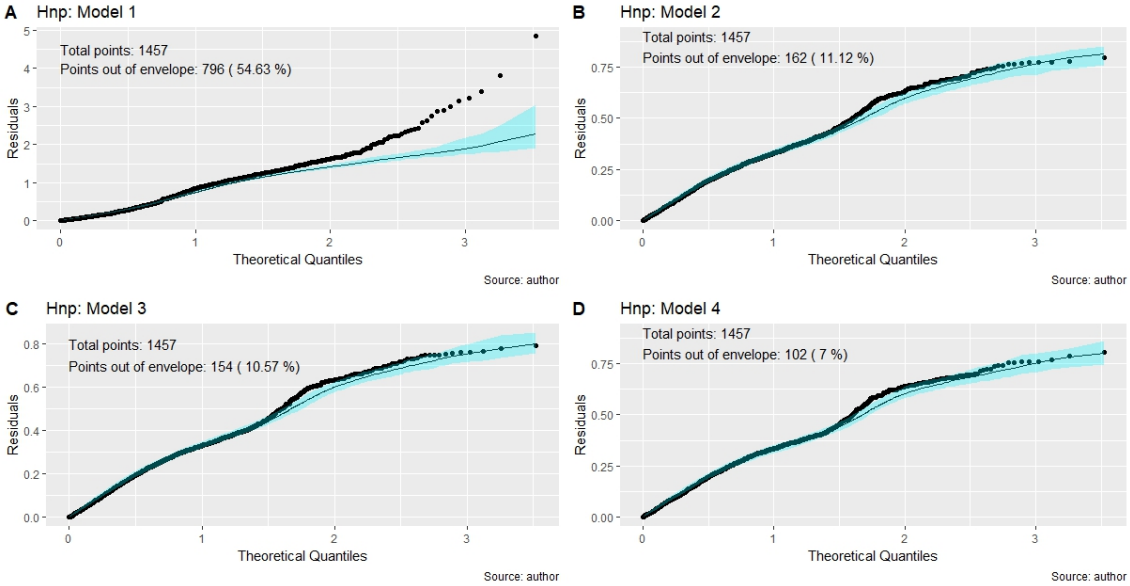

Figure 3.5. Embryo Production Data: Half normal plot with simulation envelope to A) M1 model; B) M2 model; C) M3 model; D) M4 model

In this context, the M4 model was chosen because it presented the better goodness of fit to dataset. Moreover, the variance component estimated by the proposed model, due to the random effect of crossing, suggested that there is a greater expression of genetic variance in this adjustment when compared to the other combined models, indicating a decrease in the residual variability. Thus, follows in the Figure 3.6, the residual analysis for the M4 model, in which the assumption of normality was verified, that is, the proposed adjustment is adequate. Additionally, in the Figure 3.7 we have the graphical representation of the predicted values versus observed values, considering the significant covariates.

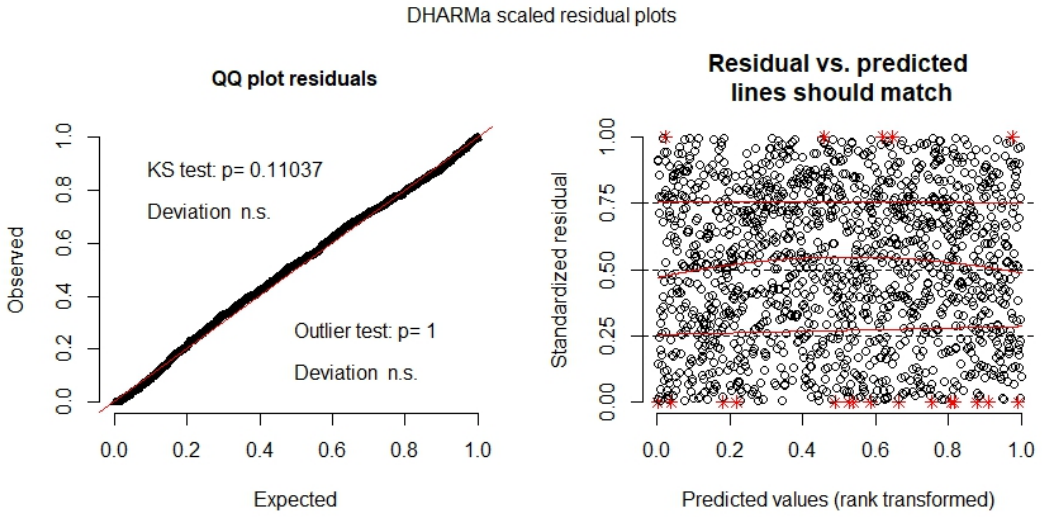

Figure 3.6. Embryo Production Data: Graph of adjusted values versus observed values; and predicted values versus standardized residuals, respectively

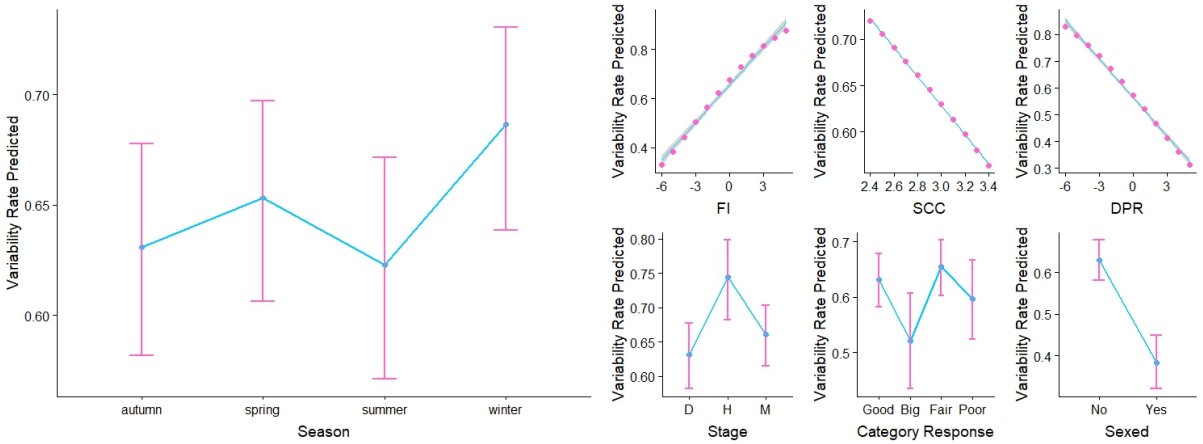

Figure 3.7. Embryo Production Data: Graphs of predicted values versus observed values for the M4 model

3.4 Discussion

As suggested by Taneja et al. (2000), the superovulation in cattle consists of the viability of a female to produce larger numbers of progeny during their reproductive life. Although it is a very useful technique for the better use genetic of the herd, there are some factors that interfere in the viable embryos number obtained in the collection process.

For understand the possible influences on the production of embryos (subsection 3.2.1) the response variable was the ratio between viable embryos number and the total embryos produced number, characterizing the embryo viability rate.

Studies of correlations between characteristics of interest provide information that may assist professionals in the efficiency of the embryo production protocol. However, the interpretations through the coefficients of simple correlations, also known as zero order, should be cautious, since the values obtained do not represent a real measure of cause and effect. The degree of association presented between two characteristics is being influenced in the presence of one third, or group of factors, causing big mistakes. Thus, to measure the dependence, or lack thereof, between two variables in a way that does not have external influences, we use the partial correlation which is a more informative measure regarding the real relations (Gujarati e Porter, 2011).

In this context, to analyze the coefficients of the simple correlations among the quantitative variables, described in subsection 3.2.4, we have a negative relationship between the variables fertility index (FI) and somatic cell score (SCC), indicating evidence that in the presence of covariables age, total performance index (TPI), milk production (Milk), and dairy pregnancy rate (DPR), all related to the donor cow, as the SCC values decrease, the FI values are increased. However, the partial correlation among these characteristics, in fact, showed a positive relationship, that is, as the FI increases, there is elevation in the SCC. This relationship was not expected because the infection of the mammary gland, or other tissues, has a negative effect on reproductive efficiency (Ribeiro et al., 2016). Therefore, for cows with high FI values was expected that they would present low SCC, because this would be a marker of breast health and

resistance to infections. One possible explanation for the result is that presenting high fertility indexes, the cow will produce larger numbers of somatic cells, in which it will be more prone to acquire infections in the mammary glands (mastitis) caused by some microorganism (Machado et al., 2000; Langoni et al., 2011).

Results of partial correlations, such as positive relation, of strong intensity between DPR with FI; and TPI with SCC is expected, since the daughter pregnancy rate is 64 % of the fertility index values, and the somatic cell score, from the cell count, comprises 5% of the TPI values (H.A, 2017). The association between TPI and SCC, in practice, indicates evidence that as the degree of infection of the mammary gland increases, there is a reduction in the total performance index of the donor cow.

Initially, the adjustment of the binomial model to data was performed only with the fixed effects variables, and the results were not approached in this work. This distribution presupposes the occurrence of constant success (Myers e Montgomery, 2010).However, the estimation of the dispersion parameter was $\hat{\phi} = 4.31$, being greater than specified by the model ($\phi = 1$), showing an overdispersion present in the data set (Hinde e Demétrio, 1998). The occurrence of the extra variation is due to variability in the success rate since the obtainment of viable embryos is influenced both genetic features related to the crossing (donor cow $\times$ donor sire) and environmental factors.

Models based on the logistic link function were proposed in different scenarios (Table 3.1). Due to the sources of variations and correlations present between the observational units, we have included the random effect at crossover level in the linear predictor, since besides assuming that these combinations come from a population with several possibilities, we can consider the genetic variability present in each individual. The authors Breslow e Clayton (1993), the authors said that not to include random effects in the linear predictor when hierarchical structures are present (the data have some nesting structure (Agresti, 2018)), may result in bad fit of the model, and interpretations of the significance of the effects erroneously.

However, the goodness of fit measures calculated under the M1 model, indicated with 5% of significance that the hierarchical model is inadequate, that is, the inclusion of the random effect only in the linear predictor was not sufficient to capture the extra variability contained in the data.

Maintaining with the linear predictor structure given in M1, we proposed the M2 model, which incorporated a random effect multiplicative in the distribution average which will accommodate overdispersion. According Molenberghs et al. (2010, 2017),these models characterized by combined will promote the adjustment of the data when there were two sources of variations, simultaneously. The absence of significance among the covariates of the model are caused due to the multicollinearity among the characteristics. Thus, some effects are being canceled in the presence of other (Gujarati e Porter, 2011).

In this way, extensions of the combined model M2 have been suggested, and the results were satisfactory. In the evaluations between the M3 and M4 models, the estimated fixed effects and predicted random effects values indicated that both adjustments behaved similarly. Equivalence was also observed for the comparison criteria, such as AIC and deviance, differing in decimal magnitude. However, when verifying the goodness of fit through of the half normal

probability with simulated envelope, it is assumed that the M4 model was the most appropriate, since, besides presenting fewer points out of the envelope, it can be said that the simplest model is as efficient as those that have more parameters. Highlights that, in general, the variables that were significant in the M3 model remained significant in M4, except for fixed effects of year and milk production, indicating in this case, evidence that the embryo viability rate remained constant over time, and that the genetic marker for milk production did not affect the number of viable embryos.

Evaluations of the reproductive efficiency offer advantages in the selection criteria of the animals that will integrate into the superovulation protocol. The significance of the estimated effects of the chosen model M4, provided important information. One of them would be the influence observed in the variable response, regarding the fertility index of the donor cow. Probably, the success of this index in embryo viability is explained by the fact that several components of reproductive efficiency was combined in a single index, such as the female ability to conceive when heifer, conceive when cow, and the daughters conception of cows, without mentioning general ability to return to cycling (show estrus) and maintain pregnancy (H.A, 2017). Thus, at the 5 % significance level, there is evidence that all these factors would be responsible for the positive result, indicating that the higher this index, the greater the number of viable embryos produced. However, when evaluating the daughter pregnancy rate (DPR), singly, this contributed negatively to the response variable. This result shows that, for the selection of embryo donors, it is not necessary to consider isolated genetic characteristics, since the positive effect of FI on the production of viable embryos has been demonstrated.

The stage of animal development involves several moments during the life of the female. For example, heifers are those that have not yet calved their calves. Lactating cows are those that are producing milk at the time of collection of embryos. Unlike the dry cows, which have already procreated, but they are not producing milk. In this context, when analyzing the variable life stage of the donor, it can be said that there were indications of a decrease in the embryo viability rate when using dry and lactating cows. These results were corroborated by Demétrio et al. (2007); Sartori et al. (2009). One possible explanation for this decrease is the fact that milk production interferes the reproductive physiology, resulting in low fertility (Wiltbank et al., 2006; Andreu-Vázquez et al., 2012).

Another possible interference factor in the production of viable embryos would be the season. During the warmer months, as in the summer, the fertility rates of the animals are lower. This happens due to the thermal stress in the animal, inducing it to a lower quality of the oocytes, and, consequently, interfering negatively in the embryo production (Thatcher et al., 2001; Sartori et al., 2002a; Hansen, 2009). We can be observed in the model chosen that the summer and autumn seasons have been negative in the response. Whereas in the spring this does not occur. Furthermore, there was significant evidence of better results when performing embryo production in winter or spring, being these seasons did not differ statistically.

The somatic cell score allows an evaluation of the mammary gland in lactating females, using as a reference the increase in the concentration of defense cells in the milk, constituting a reliable indication as to the udder infection level (Laranja e Amaro, 1998). In the M4 model, the SCC genetic trait was statistically significant in the response variable, showing evidence that

high SCC values cause a decrease in viable embryo productivity. In this context, the use of this genetic characteristic in conjunction with others to select donor cows may be effective.

One of the characteristics of great importance in artificial inseminations (AI) is the type of semen. In this process, two types of semen are used. The sexed contains only the spermatozoa with the "Y" chromosome, giving rise to the males, or only, the "X" chromosome, originating the females. And the non sexed does not change. Both types will be frozen process. However, even though there is a high number of born of the desired sex, the sexed semen has lower fertility, since it suffers separation of the spermatozoa by processes that are invasive, which can cause damages as to its viability and spermatic quality (Seidel Jr, 2003, 2007). Similarly, to those found in Seidel Jr (2007), there were indications of influence on the embryo viability rate per semen type, in which negative estimates when performing AI with the sexed type caused a reduction in the number of viable embryos.

Verification of the cow's response to the superovulation stimulus is necessary to determine whether the female was able to respond satisfactorily to the protocol. Thus, the superovulatory response in donors of the Holstein breed was significant, indicating that cows with adequate responses to superovulatory treatment produced higher estimates in the result of viable embryo production.

Based on the components of variance presented, it can be seen that the variability caused by the random crossover effect was higher in the M4 model, that is, this adjustment enables to reduce the residual variation, which was decomposed and subsequently explained by genetic variability.

In the practical context, the points that were found outside the simulation envelope in the half normal probability plot are justified by the presence of uncontrolled factors that are intrinsic to the sample. For example, donor health (Ribeiro et al., 2016), hormonal profile (Demétrio et al., 2007), and estrus expression (Pereira et al., 2016). The different ways in which the superovulation protocol is conducted by professionals should also be considered since can cause variability.

3.5 Conclusions

The embryo viability rate presented a greater dispersion than expected by the binomial model, in which this extra variation may have been caused by random uncontrolled factors. Among the proposed models, the binomial-logistic-normal-beta was satisfactorily adjusted, since the adjustment was able to accommodate both the overdispersion phenomenon and the correlation structures present in the data.

Thus, in order to achieve success in embryo production, the purpose of which is to obtain larger quantities of viable embryos, for later use in transference processes, it is important to select females with a high fertility index , since choices based on these values will provide better results in terms of efficiency in superovulation protocols.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley, New York, 3 edition.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademinai Kiado.

Andreu-Vázquez, C., Garcia-Ispierto, I., Ganau, S., Fricke, P., e López-Gatius, F. (2012). Effects of twinning on the subsequent reproductive performance and productive lifespan of high-producing dairy cows. *Theriogenology*, 78(9):2061–2070.

Bates, D., Mächler, M., Bolker, B., e Walker, S. (2015). Fitting linear mixed-effects models using lme4. R package version 3.3.1.

Bó, G. e Mapletoft, R. (2013). Strategies for donor and recipient selection, treatment, and management. *Proceedings American Embryo Transfer Association*, pages 16–24.

Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., e Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Cordeiro, G. M. e Demétrio, C. G. (2008). *Modelos lineares generalizados e extensões*, volume 33. ESALQ/USP, Piracicaba.

de Andrade Moral, R., Hinde, J., e Garcia Borges Demétrio, C. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(10).

Demétrio, C. G. B. e Zocchi, S. S. (2011). *Modelos de Regressão.* ESALQ-USP, Piracicaba, São Paulo.

Demétrio, D., Santos, R., Demétrio, C., e Vasconcelos, J. L. M. (2007). Factors affecting conception rates following artificial insemination or embryo transfer in lactating holstein cows. *Journal of Dairy Science*, 90(11):5073–5082.

Gujarati, D. N. e Porter, D. C. (2011). *Econometria Básica.* Amgh Editora, 5 edition.

H.A (2017). Updates to the Total Performance Index (TPI) and Type Composites. *Holstein Association USA*.

Hansen, P. J. (2009). Effects of heat stress on mammalian reproduction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1534):3341–3350.

Hinde, J. e Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170.

IETS (2010). *Manual of the international embryo transfer society.* Champaign (IL): International Embryo Transfer Society, 4 edition.

Kanagawa, H., Shimohira, I., e Saitoh, N. (1995). Manual of bovine embryo transfer. *Japan Livestock Technology Association*, pages 1–44.

Langoni, H., Penachio, D. d. S., Citadella, J. C., Laurino, F., Faccioli-Martins, P. Y., Lucheis, S. B., Menozzi, B. D., e Silva, A. V. d. (2011). Aspectos microbiológicos e de qualidade do leite bovino. *Pesquisa Veterinária Brasileira*, pages 1059–1065.

Laranja, L. e Amaro, F. (1998). Contagem de células somáticas–conceitos e estratégias de controle. *Rev. Balde Branco*, pages 28–34.

Laster, D. (1972). Disappearance and uptake of [125i] fsh in the rat, rabbit, ewe and cow. *Reproduction*, 30(3):407–415.

Lindsay, B. G. (1986). Exponential family mixture models with least-squares estimators. *The Annals of Statistics*, pages 124–137.

Machado, P. F., Pereira, A. R., e Sarríes, G. A. (2000). Composição do leite de tanques de rebanhos brasileiros distribuídos segundo sua contagem de células somáticas. *Revista Brasileira de Zootecnia*, 29(6):1883–1886.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime data analysis*, 13(4):513–531.

Molenberghs, G., Verbeke, G., e Demétrio, C. G. (2017). Hierarchical models with normal and conjugate random effects: a review. *SORT-Statistics and Operations Research Transactions*, 1(2):191–254.

Molenberghs, G., Verbeke, G., Demétrio, C. G., Vieira, A. M., et al. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical science*, 25(3):325–347.

Molenberghs, G., Verbeke, G., Iddi, S., e Demétrio, C. G. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111:94–109.

Myers, R. e Montgomery, D. C. e Vining, G. G. (2010). *Generalized Linear Models: With Applications in Engineering and the Sciences*. John Wiley, New Jersey, 2 edition.

Nelder, J. e Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A*, 135:370–384.

Pereira, M., Wiltbank, M., e Vasconcelos, J. (2016). Expression of estrus improves fertility and decreases pregnancy losses in lactating dairy cows that receive artificial insemination or embryo transfer. *Journal of dairy science*, 99(3):2237–2247.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, E., Gomes, G., Greco, L., Cerri, R., Vieira-Neto, A., Monteiro Jr, P., Lima, F., Bisinotto, R., Thatcher, W., e Santos, J. (2016). Carryover effect of postpartum inflammatory diseases on developmental biology and fertility in lactating dairy cows. *Journal of dairy science*, 99(3):2201–2220.

Rizzato, F. B. (2011). *Modelos para análise de dados discretos longitudinais com superdispersão*. PhD thesis, Universidade de São Paulo.

Sartori, R., Bastos, M. R., e Wiltbank, M. C. (2009). Factors affecting fertilisation and early embryo quality in single-and superovulated dairy cattle. *Reproduction, Fertility and Development*, 22(1):151–158.

Sartori, R., Sartor-Bergfelt, R., Mertens, S., Guenther, J., Parrish, J., e Wiltbank, M. (2002a). Fertilization and early embryonic development in heifers and lactating cows in summer and lactating and dry cows in winter. *Journal of dairy science*, 85(11):2803–2812.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Seidel Jr, G. (2003). Economics of selecting for sex: the most important genetic trait. *Theriogenology*, 59(2):585–598.

Seidel Jr, G. (2007). Overview of sexing sperm. *Theriogenology*, 68(3):443–446.

Self, S. G. e Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

Taneja, M., Bols, P. E., de Velde, A. V., Ju, J.-C., Schreiber, D., Tripp, M. W., Levine, H., Echelard, Y., Riesen, J., e Yang, X. (2000). Developmental competence of juvenile calf oocytes in vitro and in vivo: influence of donor animal variation and repeated gonadotropin stimulation. *Biology of reproduction*, 62(1):206–213.

Thatcher, W., Guzeloglu, A., Mattos, R., Binelli, M., Hansen, T., e Pru, J. (2001). Uterine-conceptus interactions and reproductive failure in cattle. *Theriogenology*, 56(9):1435–1450.

Thompson, J., Allen, N., McGowan, L., Bell, A., Lambert, M., e Tervit, H. (1998). Effect of delayed supplementation of fetal calf serum to culture medium on bovine embryo development in vitro and following transfer. *Theriogenology*, 49(6):1239–1249.

Verbeke e Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied statistics*, 31(2):144–148.

Wiltbank, M., Lopez, H., Sartori, R., Sangsritavong, S., e Gümen, A. (2006). Changes in reproductive physiology of lactating dairy cows due to elevated steroid metabolism. *Theriogenology*, 65(1):17–29.

Wright, J. M. (1981). Non-surgical embryo transfer in cattle embryo-recipient interactions. *Theriogenology*, 15(1):43–56.

4 A MACHINE LEARNING APPROACH TO EMBRYO TRANSFER DATA

Abstract

The ability to predict whether an artificial insemination procedure would result in pregnancy is of great value in optimizing the breeding strategies employed by dairy producers. In the past few decades, machine learning algorithms have successfully been applied for prediction tasks in a variety of fields, including animal health and productivity, being therefore very promising candidates to be used in cattle breeding as well. In this study we have applied several machine learning algorithms with the purpose of evaluating the performance of those techniques and ascertain whether it's possible to predict whether a given insemination event resulted in pregnancy in dairy cows of the Holstein breed. The resulted analyses indicated that, when applied to this particular data set, this methodology wasn't able to produce satisfactory results with regard to the algorithm performance, since the available data lacked enough observations to learn the negative class. Besides that, a comparison between classical statistical methods and these machine learning techniques was carried out, where it was found that, given the issues found in the data, the use of these techniques is discouraged, being classical statistics more appropriate for the analysis of this dataset.

Key words: Binary Data; Logistic Regression; Machine Learning; Random Forest; Viable Embryo.

4.1 Introduction

Machine Learning, a branch of the field of artificial intelligence, is a form of predictive modelling using statistical models to develop predictions that has been defined as the ability of a computer to learn without being programmed for it. It uses algorithms that analyze some "training data" first, and then generates outputs based on that analysis. After learning, predicting, find ways to improve their performance over time.

In order to understand how machine learning methods differ from the traditional workflow employed in statistics, it's useful to take a step back and think about what is a statistical problem.

In the most general sense, we are presented with some kind of natural process whose results can be measured and represented by some output variables. We also have another set of measurable quantities, our input variables, which we suspect to have some effect on our outcomes. With these in hand, there are two kinds of objectives we might be interested in achieving: advancing our understanding of how the natural process relates our inputs to our outputs, or predicting what our response is going to be under a given set of inputs. The manner in which one usually proceeds towards these goals is then to posit some data generating process, a statistical model of our original, much more complicated, natural mechanism. It is then through the study of this simpler model that information on the workings of the initial phenomenon, as well predictions of its behavior, is drawn. This is the conventional approach in statistics.

The machine learning approach differs in that it foregoes this attempt to conjure a device that can function as a drop-in replacement of the process at hand. Instead, the latter is

treated as a black box whose contents can't be known. The focus then is merely to produce a function f capable taking our inputs and predicting the outputs associated with it. While, on the surface, this might seem like a mere abdication of epistemological claims, with little impact on the statistical practitioner's work, it actually gives rise to far ranging implications on the selection of tools (algorithms) to be employed, and the evaluation of the results obtained from them (Breiman, 2001b).

While media attention has largely focused on a few selected applications of machine learning, such as its use in marketing analytics, financial forecasting, and autonomous vehicle navigation, scientific literature abounds with the used cases that seldom receive the public spotlight. Livestock breeding is one of these fields where such methods have been steadily making inroads. Aided by an accompanying set of technological advances, which enabled automatic collection of an unprecedented amount of data on animal health, fertility, and genetic characteristics, those approaches have yielded a myriad of relevant results to problems such as prediction of disease, phenotypes, insemination success, and mortality rates (Nayeri et al., 2019).

One such domain which is of particular interest here is that of examining insemination outcomes. In chapter 2, we have employed Generalized Linear Mixed Models (GLMMs) to analyse one such case. Therefore, revisiting that same data under the framework of machine learning algorithms can be useful in order to render clearer the differences in methodology and compare the results between the already established approaches and these new entrants to the statistical modelling toolbox.

Throughout this chapter, the most relevant concepts guiding predictive modelling using machine learning algorithms will be presented, drawing parallels with "traditional" statistical techniques whenever applicable. The essential theoretical background will be laid out in section 4.2, whereas in subsection 4.3.1 we will explore a practical application to the prediction of successful embryo transfer in bovines, and in section 4.4 we discuss the results obtained through the lens of prediction "strength" metrics and model interpretability.

4.2 Material and Method

4.2.1 Case Study: Embryo Transfer Data

In this chapter, we used a data set relating to embryo transfer on Holstein-breed cows, obtained from an observational study conducted at the Ruann and Maddox Dairy Farm in Riverdale, California, USA. Each data point in the final data set included 15 features and 1 binary response variable (PD) indicating whether the embryo transfer resulted in a pregnancy diagnosis ($Y = 1$) or not ($Y = 0$). The features of the data set consist of the Fertility Indices (FI) of all animals involved (donor cow, recipient cow, as well as donor sire), season of the year when the transfer took place (Season), lifecycle stage of the recipient cow (VH_LC), days from recipient cow's estrus when the transfer was made (DFE), the *corpus luteum* hemisphere where the embryo was implanted (CL_side), type of embryo (TEFF2), embryo development stage (Embryo_stage), the technician responsible for the procedure (Tech), donor cow lifecycle stage (Type), and whether the semen used was sexed or not (Sexed). Further details can be found in subsection 2.2.1.

Overall, 940 distinct combinations of donor cow and sire were performed. On Table 4.1 a summary view of the features in the data set can be found.

4.2.2 Machine Learning Algorithms

While the machine learning toolbox does borrow many techniques from the library of statistical methods, its community eventually developed a distinct vocabulary around it. The first relevant typology is that of the classes of problems where machine learning is typically employed, in this, two broad classes are of interest here:

* *Unsupervised learning*: in this setting, the main interest lies in extracting information from patterns in the data. This can be used both as an end goal in itself or as a step to simplify the data before applying other techniques.

* *Supervised learning*: in this setting, the main interest lies in *training* an algorithm to predict the response associated with a set of input variables (features). If the response variable is continuous, it is said that the task is a "regression" one, whereas if it consists of a number of discrete "classes", it is said that the task is a "classification" one (Raschka, 2015). Do note that the fact that these algorithms are usually built with prediction performance in mind doesn't mean they can't be used to obtain quality information about the problem, the opposite is actually very often the case, as shown in the examples presented by Breiman (2001b).

Since the problem at hand consists of predicting the outcome of an embryo transfer event, which can either result in success (a pregnant cow) or failure (non-pregnant), we say that this is a supervised, classification task.

No systematic approach exists to determine the most suitable machine learning algorithm to apply in a given scenario. The most common procedure is then to select a handful of suitable algorithms, train them on the data, and then compare their performance to determine the one with a better fit.

In this work, based on the binary nature of our response variable, we've chosen to use classification algorithms, such as Decision Trees, Random Forest, and Logistic Regression. In all cases, we've used the implementations contained in the popular scikit-learn package for the Python programming language (Pedregosa et al., 2011).

Decision Trees

Decision trees for classification (or simply classification trees) are built by successively partitioning the feature space in rectangular subdivisions where the predicted class is constant throughout. The splitting point s that defines a new partition is chosen by evaluating an impurity metric, such as cross-entropy or Gini index, for each valid candidate, and greedily choosing the one that maximizes impurity decrease. With this, we obtain two new subdivisions of the feature space, defined by the feature values lesser than or equal to s and those greater than it. For each one, the predicted value associated with it is defined by the majority class of training instances contained in each region (Song e Ying, 2015).

Table 4.1. Summary statistics of the data used in the study

No.	Feature	Non-Null	Type	Levels	Mean	SD	Min	25%	50%	75%	Max
1	PD	5108	Binary	2	–	–	–	–	–	–	–
2	Recipient	5108	Categorical	4070	–	–	–	–	–	–	–
3	FI_R	5108	Numeric	–	-0.46	1.49	-6.00	-1.50	-0.40	0.50	4.90
4	Season	5108	Categorical	4	–	–	–	–	–	–	–
5	VH_LC	5108	Categorical	2	–	–	–	–	–	–	–
6	DFE	5108	Categorical	4	–	–	–	–	–	–	–
7	CL_side	5108	Categorical	2	–	–	–	–	–	–	–
8	TEEF2	5108	Categorical	3	–	–	–	–	–	–	–
9	Embryo_stage	5108	Categorical	3	–	–	–	–	–	–	–
10	Tech	5108	Categorical	6	–	–	–	–	–	–	–
11	Donor_Cow	5108	Categorical	386	–	–	–	–	–	–	–
12	FI_D	5108	Numeric	–	-0.67	1.69	-5.50	-1.70	-0.70	0.50	4.00
13	Type	5108	Categorical	3	–	–	–	–	–	–	–
14	Donor_Sire	5108	Categorical	176	–	–	–	–	–	–	–
15	FI_S	5108	Numeric	–	-0.98	1.82	-4.60	-2.50	-1.10	0.40	4.50
16	Sexed	5108	Categorical	2	–	–	–	–	–	–	–

The splitting procedure outlined above is repeated until one of many stopping criteria is met. In the implementation used for this work, these include:

* A hard limit on the maximum depth the tree can reach.
* A minimum number of samples that a node eligible to further splitting has to have.
* Any further possible split would cause a decrease in impurity that is less than a defined threshold.

According to Trevor et al. (2009), some limitations of decision trees as predictors include:

* The computational cost of calculating optimal splits in high-cardinality features.
* Decision trees are inherently high-variance. Training a decision tree with the same hyperparameters on two different samples of the same population can lead to very different results.
* The prediction surface obtained by using decision trees is not smooth. This presents difficulties in applications where it's expected that the obtained function be smooth or differentiable.

Random Forest

The high-variance issue that decision trees present can be alleviated by combining multiple trees into an *ensemble*. Many such procedures exist, but of particular interest is that of *bagging*. This technique, introduced in Breiman (1996), consists of generating bootstrapped samples of the original training set, fitting a model on each sample and taking the bagged prediction to be the average over all component models' predictions. In a classification setting, the output of the bagged model is equivalent to taking the class obtained by majority voting of the ensemble members, where each individual tree has exactly one vote with equal weight.

This averaging process can be shown to reduce overall classification error as long as the original, non-bagged, classifier performs better than chance (Raschka, 2015). The result is that a bagged classifier often presents lower variance than its unbagged counterpart, but without introducing additional bias (Trevor et al., 2009).

Random forests were proposed by Breiman (2001a) as an improvement to the naïve tree bagging method. The author's main insight was that the correlation between pairs of trees in the ensemble can be reduced by constraining the feature space during the tree-building process, where, instead of considering all p features of the dataset, each tree in a random forest only has available a randomly-chosen subset of $m \leq p$ features to work with. While the degree of this decorrelation effect and its relationship with m are highly dependent on the structure of the data generating process, the intuition behind the mechanism can be summarized by the fact that a given pair of trees is less likely to agree on the classification of a given sample if the more disjoint the set of features they employ is.

This decorrelation effect helps drive the total variance of the random forest further down than a naïvely bagged tree ensemble, while still enjoying the property of not introducing additional bias into the model.

While inspecting the decision paths of 100 decision trees (that's the default number of trees in a forest under scikit-learn's implementation) is hardly approachable to a human, random forests are nevertheless still very useful for extracting information about a problem's domain. For one, remember that while easily scrutable, decision trees are inherently unstable and prone to overfitting, a problem that random forests circumvent through bagging and their clever feature subsetting scheme. But also, random forests enable one to produce a measure of the relative importance of each feature for the response prediction by summing the impurity decrease at each split a given feature is involved in over all the trees in the forest (Trevor et al., 2009).

Logistic Regression

Logistic regression is a well-attested technique in the statistician's toolbox. While gaussian linear regression works under the assumption that the response variable can be represented as a continuous value, logistic regression is a method that can be used when $Y_i \in 0, 1$, that is, it is a binary variable. In the generalized linear models, it is necessary to relate the linear predictor, which contains the independent variables, with the response variable's mean via a link function. That way, under the canonical link function given by $\theta = \log(\pi/(1-\pi)) = \eta = \mathbf{X}'\beta$, where $\log(\pi/(1-\pi))$ is the logit link, one obtains that the logistic regression model is expressed by $p(y|\mathbf{x}) = \exp(\mathbf{X}'\beta)/(1+\exp(\mathbf{X}'\beta))$ (Cordeiro e Demétrio, 2008).

In a machine learning setting, logistic regression is frequently used as a classification model. In order to use it for such ends, one must employ a decision rule that converts the probabilities output by the model into class values. In a binary classification setting, one common choice for a such a rule is to define a threshold, usually 0.5, such that $\hat{y}(x) = 1 \iff p(y = 1|x) > 0.5$ for the positive class, and otherwise $\hat{y}(x) = 0 \iff p(y = 0|x) \leq 0.5$. One shortcoming of such models, however, is that they can't model nonlinear decision boundaries, therefore it is prone to misclassification in settings where the classes are not linearly separable (Murphy, 2012).

Model Evaluation Methods

One of the most important steps in ascertaining a ML algorithm's reliability is the validation of results. Some of the metrics proposed in Raschka (2015), which are also used in this work, are:

i) Confusion Matrix: helps understand a classifier's performance. For binary classification tasks, it has four entries. The rows denote the number of samples that truly belong to each class, whereas the columns contain the number of samples that were predicted by the classifier to belong to a class. Therefore, we have in the main diagonal the number of True Negatives (TN) and True Positives (TP) respectively, whereas the off-diagonal contains the number of False Positives (FP) and False Negatives (FN).

ii) Accuracy: Is the percentage of correct predictions, calculated with $(TP+TN)/(TP+TN+FP+FN)$.

iii) Precision: Is the percentage of samples predicted to belong to the positive class which were correctly classified, with formula given by $TP/(TP + FP)$.

iv) Recall: Also known as sensitivity, is the percentage of values belonging to the positive class that were correctly classified $TP/(TP + FN)$.

v) F1: Is the harmonic mean of precision and recall, being calculated with $(2 * precision * recall)/(precision + recall)$.

vi) Receiver Operating Characteristic (ROC) Curve: is the probability curve displaying the classifier's performance, showing how true positive rate (sensitivity) behaves under changing false positive rate (inverse specificity) . Inverse specificity is given by 1 - specificity, with specificity defined by TN/(TN + FP).

vii) Area Under the ROC Curve (AUC): represents how well the algorithm is able to separate the different classes. The higher the AUC, the better the model is able to predict the classes. Therefore, a well-performing algorithm must have an AUC value close to 1.

4.3 Results

4.3.1 Data Analysis

We begin by checking the balance between the classes in the target variable. This is important, because the algorithm needs to be trained on a reasonable amount of samples for each class in order to learn the relationships between them and the input variables. The relevant literature suggests that the minority class should comprise at least 10% of the training data, otherwise additional processing might be necessary in order to fit a model with adequate performance (He e Garcia, 2009). In Figure 4.1 we can observe that there's no significant unbalancing between classes, with the majority class representing about 62% of the samples. Therefore, we can conclude that the data set contains a healthy amount of samples for both the success and failure classes and there's no need to apply techniques specific to unbalanced data.

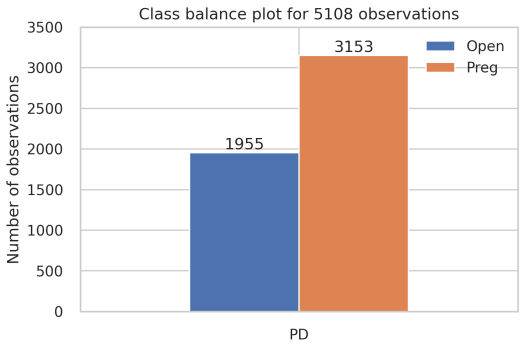

Figure 4.1. Embryo Transfer Data: Bar plot illustrating the balance between classes in the target variable

Regarding our numerical features, we can see in Figure 4.2 a pair plot relating the fertility indices of the donor and recipient cows, as well as the donor sire. On the main diagonal,

the histograms for each fertility index, colored by pregnancy success, can be seen, whereas the other elements contain pairwise scatterplot relating the fertility indices of different animals. From this, we can draw that no fertility index alone presents enough information to adequately distinguish between the two classes, as the histograms for both are completely overlapping, but also that the inspection of pairwise interactions between them similarly displays no obvious structure which could be exploited to classify our observations.

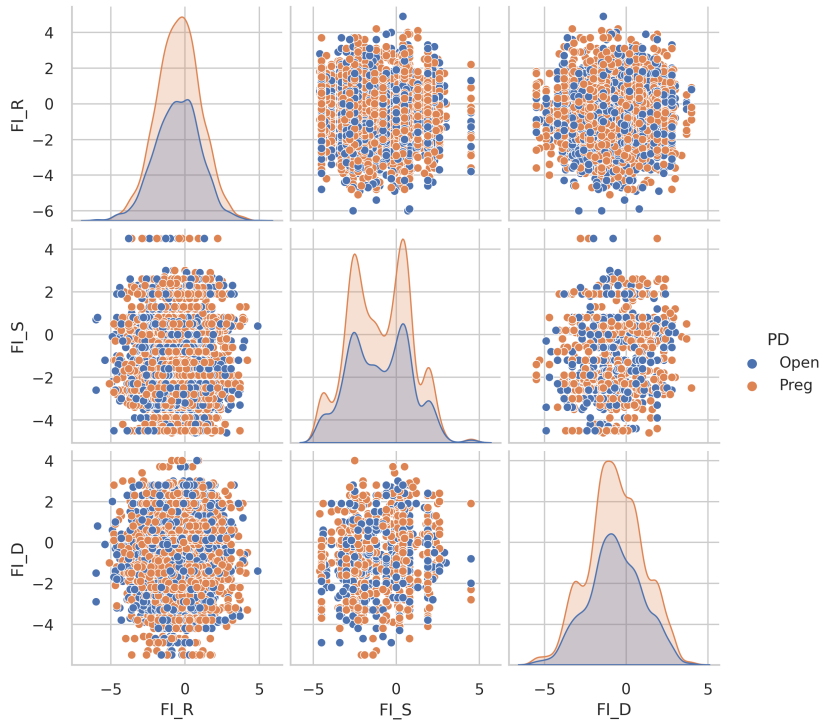

Figure 4.2. Embryo Transfer Data: Scatterplot matrix for the continuous features with regards to the target variable (open/pregnant status)

On the other hand, in Figure 4.3 we can observe the pairwise relationships between different categorical variables with regards to pregnancy success rates. It should be noted that, while some pairings are associated with success rates of upwards to 75% (*ie.* highly informative of pregnancy success), none have scores significantly lower than 50% (*ie.* highly informative of pregnancy failure). This means that while some combinations could be exploited by an algorithm to learn what factors contribute to positive conception outcomes, when it comes to learning combinations of factors that lead to negative outcomes, it would have to rely on combinations that are very weakly associated with them at best, or essentially behave like noise at worst (*ie.* that have very low information content pointing towards either success or failure). We believe this state of affairs to be one of the reasons why we stumbled upon the problems to be discussed in section Model Results.

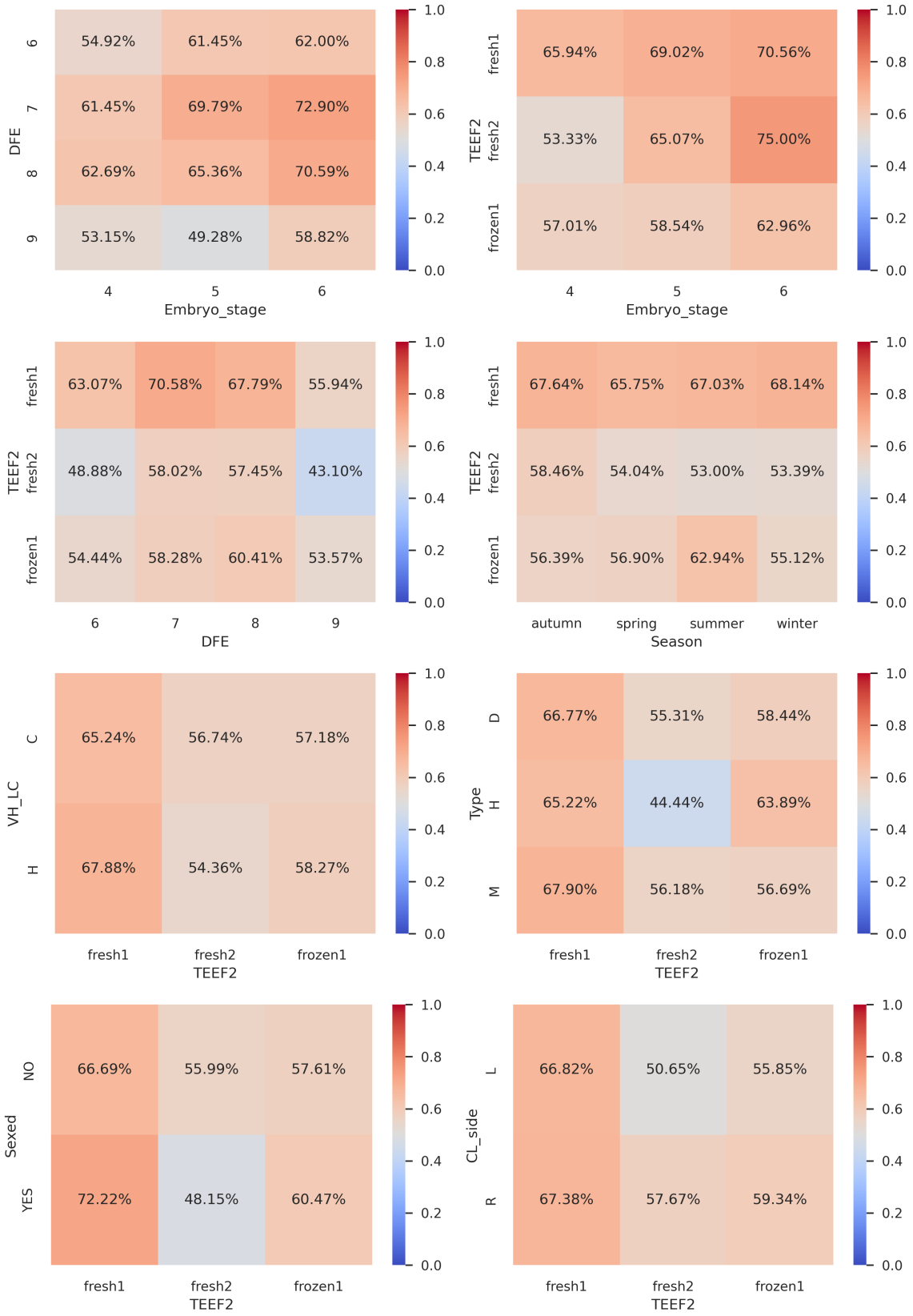

Figure 4.3. Embryo Transfer Data: Pregnancy success rates for pairwise combinations of categorical variables levels

4.3.2 Model Results

For the evaluation of algorithm performance, the dataset was split into a training and a validation set. Two split proportions were tested, the first using an 80:20 training to validation instances ratio, the second using a 70:30 one. For each splitting scheme, the same procedure two-step was carried out for all models, it consisted in first using the training set to perform a 10-fold cross validation, then fitting the model against the whole training set and assessing its performance using the validation one. The number of folds was chosen to be 10 because it's the recommended value on the relevant literature (Raschka, 2015).

The mean AUC scores and their standard deviations, as obtained from the cross-validation procedure from all available features, except for the donor sire's fertility index, which has been discarded for the same reasons as presented in chapter 2, as well as the results for the test set can be seen in Table 4.2. The AUC of predicting conception outcomes for all the models was the highest for Random Forest and Logistic Regression. Furthermore, when comparing the two dataset splitting schemes, it can be observed that both behave in a similar manner, differing only in slight AUC performance fluctuations. For the purposes of interpreting whether a given AUC value is acceptable, Hempstalk et al. (2015) suggests considering a number between 0.5 and 0.75 to be fair, and above this to be a good fit.

Table 4.2. Summarized results for 10-fold cross validation applied on the training set for different algorithms, including the AUC's mean value ("CV-Fold") and its standard deviation ("SD") over all folds. The values under "Test" correspond to the validation carried out on the hold-out test data, considering all features

	80:20 Split			70:30 Split		
Algorithm	CV-Fold	SD	Test	CV-Fold	SD	Test
Decision Tree	0.51*	0.02	0.52	0.50	0.02	0.52[a]
Random Forest	0.59*	0.03	0.56[a]	0.58*	0.03	0.56
Logistic Regression	0.57*	0.03	0.56	0.58*	0.03	0.55[a]

* p-value ≤ 0.05: t-test significance of difference from an AUC of 0.5

[a] p-value ≤ 0.05: t-test significance of difference between CV-Fold and Test-set AUC scores.

When it comes to the significance of the results obtained, the Decision Tree model was the only one to fail, in the 70:30 scenario, in obtaining a cross-validation AUC score significantly different from 0.5 (baseline), the value which characterizes a random-guessing classifier. When comparing the CV-Fold results to those on the test set, it was found that the models exhibited alternating results depending on the proportion of samples used for training, with the Decision Tree and Logistic Regression displaying consistent performance for the 80:20 split, whereas Random Forest was the one to achieve this in the 70:30 experiment. At any rate, the fact that Logistic Regression and Random Forest were top performers in predicting conception outcomes had been previously observed in Hempstalk et al. (2015), where the difficulties in improving the results obtained from the Random Forest were attributed to the high number of non-informative features in one of their datasets. In the case of the dataset used in this study, we have already

highlighted in subsection 4.3.1 how many pairwise combinations of categorical features, such as the type of embryo versus season of the year, exhibited similar issues, being particularly noninformative of failures in obtaining a pregnancy.

For comparative ends, the reduced models proposed in chapter 2 will be tested. Besides that, since there was no pronounced difference in performance between the two splitting strategies, only the 80:20 split will be used in the next sections. The features used for each model can be seen in Table 4.3.

Table 4.3. Features considered for each of the reduced models

Model	Features considered
Reduced 1	Type, CL_side, FI_R, TEEF2, Embryo_stage, DFE
Reduced 2	FI_R, Type, CL_side

On Figure 4.4, we have plotted the ROC curves obtained for both reduced models, as well as the one built using all available features (Full set). As can be seen, the "Reduced 1" set of features presents a slight performance improvement across all the range for Logistic Regression, as well as exhibiting a better behavior around the 0.2-0.5 false positive rate range for Random Forest. Meanwhile, the "Reduced 2" set performed the poorest, overlapping over several points with the "Baseline", which denotes the line of no discrimination, representing the performance of a classifier that operates by random guessing.

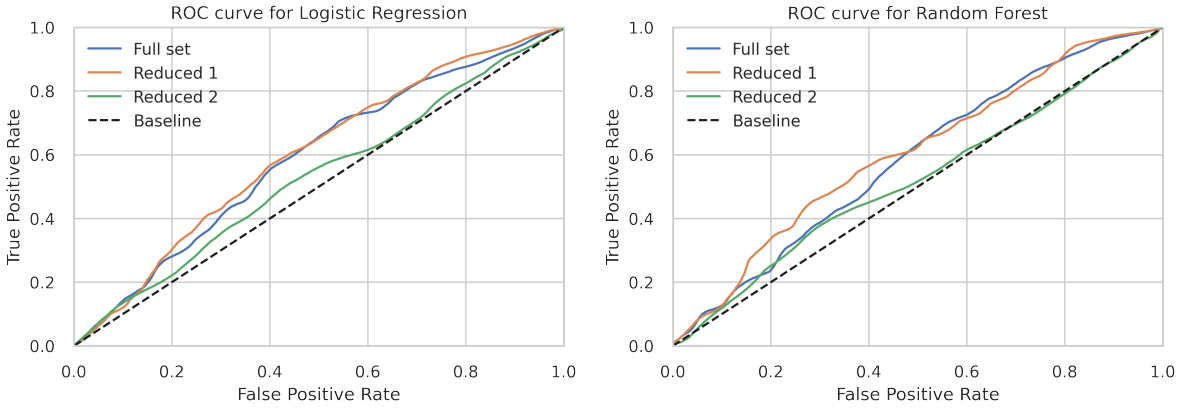

Figure 4.4. Embryo Transfer Data: comparison of ROC curves for Logistic Regression and Random Forest, respectively, using different feature subsets

Apart from the ROC curve charts, we have also compiled several other classification performance indicators for each of the models, which can be seen in Table 4.4. These help putting in perspective some of the behaviors observed on the ROC curves, but can also highlight the perils of adopting any one scalar metric as the sole indicator of an algorithm's performance. For instance, the models trained with the "Reduced 2" set had metrics either only slightly worse, or, in some cases, even better than those obtained from the other feature sets. Despite that, it has been seen that they had performed extremely poorly on the ROC curve.

Moreover, the discrepancy between precision and recall for the Random Forest algorithm indicated that it has "learnt" to classify samples as disproportionately belonging to the positive class. This behavior means that it is successfully able to guess when an insemination event resulted in pregnancy, which accounts for the high recall, but at the cost of incurring too many false positive errors, which in turn lead to the relatively low precision score. This inability in capturing the patterns indicative of a conception failure leads us to, once again, suspect the lack of such information in our dataset, as first mentioned in subsection 4.3.1, has hindered the model training process.

Table 4.4. Performance metrics for the models obtained using different sets of features

	Logistic Regression (LR)			Random Forest (RAF)		
Metrics	Full set	Reduced 1	Reduced 2	Full set	Reduced 1	Reduced 2
Accuracy	0.60	0.60	0.57	0.60	0.60	0.59
AUC	0.59	0.60	0.53	0.58	0.60	0.52
F1	0.67	0.67	0.70	0.75	0.75	0.74
Precision	0.66	0.65	0.60	0.60	0.60	0.60
Recall	0.69	0.69	0.84	1.00	1.00	0.99

The only metric that has consistently replicated the behavior seen in the analysis of the receiver operating characteristic plot was AUC, which is itself obtained by integrating over the ROC curve. This unreliability of performance metrics built from the confusion matrix, as well as the usefulness of the ROC curve as a model diagnostic tool had already been observed in Hamel (2009).

4.3.3 Ranked Predictions

It is generally expected that a model would output more extreme scores to those instances where it has a high confidence in their predicted classes. In order to study to what degree our models agree with or defy that assumption, we have ranked the records in the test set by their prediction scores for each class and evaluated the classification accuracy under several percentiles. The result of this procedure is in Table 4.5, where the 5% percentile contains the highest scores, being accumulated up until the 25%, shown separately for each class.

When observing the top rows of the table, corresponding to the observations with the highest scores towards the pregnancy success class, it can be seen that the cumulative accuracy for increasing percentiles doesn't show expressible variation. The only exceptions being the LR models for the "Full set" and "Reduced 2" feature sets, which display the aforementioned expected property of being highly assertive for the observations they attributed a high chance of success. Nevertheless, the accuracies for up to the top 25% positive class scores remain reasonably high.

Table 4.5. Prediction accuracy of each model over classification score percentiles for each class

Positive class ($Y = 1$)						
	Logistic Regression (LR)			Random Forest (RAF)		
Percentiles	Full set	Reduced 1	Reduced 2	Full set	Reduced 1	Reduced 2
5	0.69	0.59	0.67	0.65	0.71	0.60
10	0.69	0.62	0.68	0.69	0.65	0.62
15	0.65	0.66	0.68	0.67	0.66	0.63
20	0.68	0.63	0.63	0.65	0.70	0.64
25	0.66	0.64	0.62	0.64	0.71	0.64
Negative class ($Y = 0$)						
5	0.58	0.58	0.25	–	1.00	0.00
10	0.55	0.58	0.31	–	1.00	0.00
15	0.51	0.58	0.44	–	1.00	0.00
20	0.49	0.55	0.39	–	1.00	0.00
25	0.51	0.57	0.40	–	1.00	0.00

When examining the equivalent data for the negative class, however, a much more problematic behavior is seen. While the LR models displayed modest results throughout, despite not being very assertive even in the top percentiles, the RAF models produced rather inconsistent values. That can be attributed to the fact that they had very few samples predicted to have resulted in pregnancy failure. In fact, the RAF obtained from the full set predicted that not one single sample would have that result, thus the missing data for that model in the table, while the ones built using reduced 1 and 2 had, respectively, 2 and 10 samples in total predicted for that class, which explains the extreme accuracy values obtained from them. This corroborates the behavior discussed before, and further reinforces that the models aren't able to determine when a given insemination event is unlikely to result in a pregnancy for the recipient cow.

4.4 Discussion

The ability to predict the results of embryo transfer procedures by using machine learning algorithms wasn't entirely successful. The first problem was actually encountered before a single model was fit to the data, when a preliminary analysis showed that almost no combination of pairs of features revealed a pattern that could be exploited to correctly separate the two classes under study. There, it was also found that even in the cases where some information could be garnered, it was overwhelmingly in favor of predicting the positive class. While concerning, those findings weren't enough to, a priori, rule out the possibility of applying machine learning, as some of the most sophisticated algorithms are known to be capable of extracting information from interactions arising from many features, as well as, recognizing underlying patterns that are not evident under mere visual inspection of the data. In this context, random forests in particular are one such example of model where the literature largely agrees as being able to autonomously learn feature interactions (Trevor et al., 2009; Wright et al., 2016), despite

there being controversies surrounding whether these are accurately represented under common feature importance evaluation methods, which could hurt the model's overall interpretability (Wright et al., 2016).

About the experiments carried out in this study, it could be seen that varying the training to test samples ratio didn't meaningfully affect model performance. Also, when the result of the same feature selection procedure carried out on chapter 2 was used in the machine learning models, it was found that, even though the selection scheme took into account the features that would optimize the performance of a model with a very different structure, at least one of the subsets obtained from the procedure resulted in an improvement for the machine learning models as well.

The results were, similarly, not so great when the minimal set (Reduced 2) was used for either for machine learning or Generalized Linear Mixed Models (GLMM) . For the random forest models in particular, it was observed that even though they'd behave a little less conservatively, predicting more samples as belonging to the negative class, their discriminative ability as evaluated on the ROC curve suffered greatly. However, the Reduced 1 subset resulted in a performance gain both for Random Forests as well as for GLMMs. However, for the latter model, the improvement can be attributed to the fact that it has the ability to incorporate additional information as random effects in the linear predictor (Breslow e Clayton, 1993), rather than relying solely on the values of fixed effects features for a given sample, as is the case for the machine learning model employed in this work.

In order to bring clarity on the obtained results, a literature review was carried out in order to find other works which have attempted to use machine learning methods for a similar purpose. In that spirit, two other notable works were found dealing with the problem of predicting conception outcomes in bovines with machine learning, Shahinfar et al. (2014); Hempstalk et al. (2015). In these, both authors present a largely overlapping selection of ML algorithms. They differ in that the former relied on extensive datasets with over $100,000$ data points each, comprising many herds in US-based dairy farms, which used exclusively Holstein-breed cows in a year-round mating system. The latter in turn, used a dataset of approximately $6,500$ insemination records collected from Irish research farms, which employ a seasonal mating system and where Holstein-Friesian animals are predominantly, but not exclusively, used.

With this evidence in hand, we find that our results are largely in line with both. Hempstalk et al. (2015) in particular, despite obtaining models with better performance indicators, has also found that logistic regression outperformed all other algorithms, including random forest, and that in their models there was an overwhelming tendency to misclassify negative samples.

The fact, then, that logistic regression performed better for both classes, despite not being able to model nonlinear decision boundaries, reinforces the fact that there are still no definitive rules to determine beforehand which machine learning technique will better fit a given data set, or if any will fit at all. In the end, all the researcher can rely on for that is their past experiences working with similar data and performing rigorous empirical testing.

4.4.1 Statistics versus Algorithms

In the generalized linear mixed models' methodology, there is the possibility of taking into account both fixed as well as random effects. The possibility of including both of these effects in the linear predictor becomes particularly useful when dealing with binary data, given that those frequently display a variance that is much bigger than the one specified by the model, a phenomenon also known as overdispersion (Breslow e Clayton, 1993).

The application presented in chapter 2 was able to satisfactorily accommodate the dispersion present in the data set, by fitting a Bernoulli-logistic-normal model. Meanwhile, the machine learning community hasn't devoted as much attention to the potential benefits of enriching their random forest algorithm with random effects. This can be seen in the comparisons carried out throughout this work, where, even though there was a similarity between the results obtained from both methodologies, the GLMM has performed better, due to having a structure with the ability to incorporate both fixed as well as random effects.

Some approaches to take them into account do exist, and it is reported that the mixed-effects machine learning models, that is, those obtained by incorporating random effects, outperform those that only consider fixed effects (Ngufor et al., 2019). However, there are other studies that claim no expressive improvement could be extracted by adding such effects (Fokkema et al., 2021).

It is understood then, that there is a severe lack of understanding regarding how random effects should be correctly implemented in machine learning, besides that, how the mixed-effects models obtained should be benchmarked, and the exact scenarios where they should be employed to extract a tangible benefit, evidencing, in this case, that this idea remains an unconsolidated field, where additional research is needed in order to develop a systematic understanding of its nuances.

While the machine learning community hasn't been as enthusiastic in adapting existing algorithms to take random effects into consideration, there exist some evidence exists that not only it is possible, but also that the mixed-effects machine learning models obtained by doing it exhibit improved performance when tested in a variety of datasets (Ngufor et al., 2019).

4.5 Conclusions

The machine learning algorithms employed in this study have been found to have only moderate success at predicting the outcomes of embryo transfers. In this regard, it was found that all of these techniques used in this study were able to predict much more adequately the cases of success rather than those where the transfer didn't result in pregnancy. Furthermore, the successful application of a generalized linear mixed effects model to the same problem indicates that the addition of random effects to machine learning algorithms could bring great benefit to their performance, but the scant amount of research devoted to these models poses a significant barrier to their widespread adoption.

With this, while machine learning algorithms remain a valuable tool for data science, there still abound many applications where classical statistical models are capable of delivering better results, especially when there's not enough data for a machine learning algorithm to be

able to learn the set of behaviors that characterizes each class.

References

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001a). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Breslow, N. E. e Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Cordeiro, G. M. e Demétrio, C. G. (2008). *Modelos lineares generalizados e extensões*, volume 33. ESALQ/USP, Piracicaba.

Fokkema, M., Edbrooke-Childs, J., e Wolpert, M. (2021). Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, 31(3):329–341.

Hamel, L. (2009). Model assessment with roc curves. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pages 1316–1323. IGI Global.

He, H. e Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Hempstalk, K., McParland, S., e Berry, D. (2015). Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of dairy science*, 98(8):5262–5273.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nayeri, S., Sargolzaei, M., e Tulpan, D. (2019). A review of traditional and machine learning methods applied to animal breeding. *Animal health research reviews*, 20(1):31–46.

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., e McCoy, R. G. (2019). Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin a1c. *Journal of biomedical informatics*, 89:56–67.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Raschka, S. (2015). *Python machine learning*. Packt publishing ltd.

Shahinfar, S., Page, D., Guenther, J., Cabrera, V., Fricke, P., e Weigel, K. (2014). Prediction of insemination outcomes in holstein dairy cattle using alternative machine learning algorithms. *Journal of dairy science*, 97(2):731–742.

Song, Y.-Y. e Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130.

Trevor, H., Robert, T., e Jerome, F. (2009). The elements of statistical learning: data mining, inference, and prediction.

Wright, M. N., Ziegler, A., e König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC bioinformatics*, 17(1):1–10.

5 FINAL CONSIDERATIONS

The work exposed some extensions of the Generalized Linear Models, with application in zootechnical data. In the analyses performed in chapter 2, from the study of embryo transfer, and the pregnancy response variable was evaluated. It was concluded that there was a lack of adjustment by the Bernoulli model, and it was explained by the possible absence of variables in the linear predictor. However, it can be said that when adjusting the Bernoulli-logistic-normal, we have satisfactory results were obtained in the modeling.

In chapter 3, whose variable response was embryo viability, from the study of embryo production, a feature present was the phenomenon of overdispersion, which was adjusted through the binomial-logistic-normal-beta model.

For chapter 4, we have revisited the embryo transfer data used in chapter 2 now using the tools of machine learning. There, a comparison was made, where the results weren't entirely satisfactory, due to the methodology being use not being able to accommodate all the particularities of the data set.

Another conclusion was that in the process of reproductive biotechnology the phase related to the production of embryos, the indices of the donor cows were important to obtain better results regarding the embryo viability rate. On the other hand, in the transfer phase, only the recipient cow indexes were significant for success in the pregnancy. Moreover, there were disagreements regarding the seasons, since for the production of embryos the factor influenced the response variable. Whereas, in the embryo transfer process the pregnancy condition was not altered by the time.

In general, for future works, studies related to variance and covariance structures should be approached in order to better understand the possible dependencies among the covariates of the model. In addition, simulation studies will be necessary to understand the proposed methodologies. Furthermore, when it comes to the machine learning techniques, we believe an in-depth study could be carried out on the possibility of including random effects in order to extract better results than the ones obtained in this work.

APPENDIX

Appendix A - R Script

Verifying the effects of fertility indices - donor and recipient

```
#########################################################################
vh_lc_avg <- dadosTransform %>%
group_by(VH_LC) %>%
dplyr::summarise(avg_pd = mean(PD))

receptora <- ggplot(dadosTransform, aes(x = FI_R, y = PD)) +
geom_point() +
geom_smooth(method = "loess", span = 0.5, se=FALSE) +
geom_hline(aes(yintercept=avg_pd), vh_lc_avg, color="red") +
labs(x = "Recipient Fertility Indexes ",
y = "Probability of Pregnancy") +
facet_wrap(~ VH_LC)

#########################################################################
type_avg <- dadosTransform %>%
group_by(Type) %>%
dplyr::summarise(avg_pd = mean(PD))

doadora <- ggplot(dadosTransform, aes(x = FI_D, y = PD)) +
geom_point() +
geom_smooth(method = "loess", span = 0.5, se=FALSE) +
geom_hline(aes(yintercept=avg_pd), type_avg, color="red") +
labs(x = "Donor Fertility Indexes ",
y = "Probability of Pregnancy") +
facet_wrap(~ Type)
```

Models considered in the analyses - Chapter 2

```
#########################################################################
# Fixed effects model
Mod1 <- glm(cbind(PD, 1-PD) ~  VH_LC + CL_side + FI_D + FI_R +
Season + Tech + Sexed + Type + TEEF2 + Embryo_stage + DFE +
Season:VH_LC:TEEF2 + TEEF2:Embryo_stage:DFE,
family = binomial, data = dadosTransform)

# Significance analysis
anova(Mod1, test = "Chisq")
```

```
summary(Mod1)

# Model selection
AIC(Mod1)
BIC(Mod1)
deviance(Mod1)
-2*(logLik(Mod1))
Anova(Mod1)
drop1(Mod1, test="Chisq")

#########################################################################
# Random and fixed effects models
Mod2 <- glmer(cbind(PD, 1-PD) ~ Season + Tech + Sexed + Type + VH_LC +
FI_D + FI_R + TEEF2 + Embryo_stage + CL_side + DFE + Season:VH_LC:TEEF2 +
TEEF2:Embryo_stage:DFE + (1|Donor_Cow) + (1|Recipient) + (1|Donor_Sire),
family = binomial,
data = dadosTransform,
glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e4), boundary.tol = 1e-3))

# Significance analysis
anova(Mod2, test = "Chisq")
summary(Mod2)

# Model selection
AIC(Mod2)
BIC(Mod2)
-2*(logLik(Mod2))
Anova(Mod2)
drop1(Mod1, test="Chisq")
```

Models considered in the analyses - Chapter 3

```
#########################################################################
Mod1<- glmer (resp ~ FI + SCC + DPR + TPI + MILK + Age + Time +
   Season + Type + Type_2 + SEXED + (1|Donor_Sire) + (1|Donor_Cow),
   family = binomial (link=logit),
   data = dados,
   glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e4), boundary.tol = 1e-3))

# Significance analysis
summary(Mod1)

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

```
Mod2 <- glmmTMB(resp ~ FI + SCC + DPR + TPI + MILK + Age + Time +
Season + Type + Type_2 + SEXED + (1|Donor_Sire) + (1|Donor_Cow),
family=list(family="betabinomial",link="logit"),
data = dados,
control = glmmTMBControl(optCtrl = list(iter.max=1e3, eval.max=1e3)))
```

Appendix B - Python Script

```
#########################################################################
# Defining the variables: target and features

target = 'PD'

features_full_geral = ['Tech', 'VH_LC', 'CL_side', 'Embryo_stage', 'Sexed',

"Season", "Type", "FI_R", "FI_D", "DFE", "TEEF2"]

features_full = ['Embryo_stage',"Type", "FI_R", "FI_D", "FI_S", "DFE", "TEEF2"]

features_redux_1 = ['Type', 'CL_side',  "FI_R", "TEEF2", "Embryo_stage", "DFE"]

features_redux_2 = ['FI_R', 'Type', "CL_side"]
```

Constructing the proposed algorithms - Chapter 4

```
#########################################################################
# Train-test: splitting 80-20
## For 70-30 this process is similar

dfTrain, dfTest = train_test_split(dados[features_full + [target]], test_size=0.2,
random_state=20220614)

dfTest.PD[dfTest.PD == 1].count()

#########################################################################
# Encoding categorical features

dfTrainDummies = pd.get_dummies(dfTrain, drop_first=True)

dummy_features_full = dfTrainDummies.drop(columns="PD").columns

dummy_features_redux_1 = [feat for feat in dummy_features_full
```

```
if (feat in features_redux_1)
or (feat.split("_")[0] in features_redux_1)
or ("_".join(feat.split("_")[0:2]) in features_redux_1)]

dummy_features_redux_2 = [feat for feat in dummy_features_full
if (feat in features_redux_2)
or (feat.split("_")[0] in features_redux_2)
or ("_".join(feat.split("_")[0:2]) in features_redux_2)]

#########################################################################
feature_sets = models = {
"full":    dummy_features_full,
"redux_1": dummy_features_redux_1,
"redux_2": dummy_features_redux_2
}

#########################################################################
# setting proposed algorithms

models = {
"DT": DecisionTreeClassifier(),
"RAF": RandomForestClassifier(),
"LR": LogisticRegression()
}

#########################################################################
# fitting models

scaler = StandardScaler()

training_scalers = {k: StandardScaler().fit(dfTrainDummies[v])
for k, v in training_sets.items()}

scaled_training_sets = {k: v.transform(dfTrainDummies[training_sets[k]])
for k, v in training_scalers.items()}

#########################################################################
df_features = scaler.transform(dfTrainDummies[feature_sets["full"]])

df_response = dfTrain[[target]][target]

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

```
for model in models.values():
model.fit(df_features, df_response)
```

Cross Validation

```
########################################################################
def evalModel(model, X, Y):

Kfold = StratifiedKFold(n_splits = 10, shuffle = True)

cv_results = cross_val_score(model, X, Y, cv = Kfold, scoring = "roc_auc",
n_jobs = 8)

return cv_results

########################################################################
# Result: validation function's output when applied to the proposed models

results = {name: evalModel(model, df_features, df_response) for (name, model)
in models.items()}

########################################################################
# Baseline: simplest model

df_response.sum() / df_response.shape[0]

########################################################################
# Training metric for the algorithms

for name, result in results.items():
print(f"{name}: {result.mean()} +- {result.std()}")

########################################################################
# Box plot for cross validation on the training set

boxResults = pd.DataFrame(results).melt()
(
ggplot(boxResults, aes("variable", "value"))
+ geom_boxplot(color = "#1F3552", fill="#4271AE", alpha=0.7, outlier_shape =".",
outlier_color="steelblue")
+ xlab("Models")
+ ylab("AUC")
+ ggtitle("Algorithm comparison")
```

```
)

#########################################################################
# ROC e AUC

roc_viz = ROCAUC(models[benchModel])

roc_viz.fit(df_features, df_response)

roc_viz.score(dfTestFeats, dfTest[target])

roc_viz.poof()
```