Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Modelo de regressão logística para a predição de estado de saúde animal a partir de variáveis escalares e funcionais

---

Relatório Científico Final do projeto na modalidade Iniciação Científica, fomentado pela Fundação de Amparo à Pesquisa do Estado de São Paulo.

Projeto FAPESP: 2020/01436-0

Pesquisador Responsável:
Rodrigo Forti

Americana
2021

Informações Gerais do Projeto

- Título do projeto: Modelo de regressão logística para a predição de estado de saúde animal a partir de variáveis escalares e funcionais
- Nome do pesquisador responsável: Rodrigo Forti
- Instituição sede do projeto: Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas
- Equipe de pesquisa: Rodrigo Forti e Mariana Rodrigues Motta
- Número do projeto de pesquisa: 2020/01436-0
- Período de vigência: 01/09/2020 a 31/08/2021
- Período coberto por este relatório científico: 11/02/2021 a 31/08/2021

Conteúdo

Resumo

Introdução

O período de transição entre o final da gravidez em vacas e o período de lactação precoce (também chamado período periparturiente) é certamente o estágio mais importante do ciclo de lactação destes animais. A maioria das doenças infecciosas e distúrbios metabólicos ocorre durante esse período. A febre do leite, a cetose, as membranas fetais retidas, a metrite e o abomaso deslocado afetam as vacas diminuindo a produção de leite e prejudicando o seu bem-estar durante o período peripartidário. Há uma maior suscetibilidade para doenças durante o período peri-parturiente que leva, por exemplo, ao aumento de incidência de mastite no animais. Assim, a ocorrência de problemas de saúde se concentra proporcionalmente no período periparturiente, que é relativamente curto, o que certamente preocupa produtores de leite (Drackley ,1999). Como afirmado por Goff and Horst (1997), a transição do estado gestante, não lactante, para o estado não gestante e lactante, é muitas vezes uma experiência desastrosa para a vaca. Alguns problemas de saúde e reprodutivos podem ser resultado do aumento do estresse de que vacas de alta produção estão sob no início da lactação (Kaufman et al. 2018).

Além disso, no início da lactação, a ingestão alimentar é incapaz de atender às demandas de alta produção de leite. A vaca, portanto entra em um período de balanço energético negativo que leva à mobilização de reservas corporais para equilibrar déficit entre consumo de energia dos alimentos e produção de energia do leite (Lago et al. 2001). Lago et al. (2001) também defendem que processo de mobilização parece afetar o bem-estar da vaca e outras vias biológicas comprometidos à medida que a energia de entrada é direcionada para produção.

Assim, a saúde do animal durante o período de lactação é fator determinante na lucratividade dos produtores. Limitações nutricionais ou de manejo durante esse período podem impedir a capacidade da vaca de atingir a produção máxima de leite. Desta forma monitorar índices vitais da vaca, assim como do ambiente onde está inserido, é condição essencial para aprimorar o bem-estar animal e a lucratividade no período de alta incidência de doenças em vacas periparturientes

Dados

Os dados a serem utilizados neste projeto fazem parte de uma base de dados do laboratório Dairy Cattle Biology and Management da Universidade de Cornell, EUA. Estes dados foram intermediados pelo pesquisador Guilherme Rosa, da University of Wisconsin, EUA, mediante colaboração no projeto FAPESP 2017/15306-9, sob coordenação da Profa. Dra Nancy Lopes Garcia, e tendo como no projeto a Profa. Dra. Mariana Rodrigues Motta.

Os dados foram coletados a partir da análise clínica diária do estado de saúde 500 vacas. Os dados contém informações sobre características individuais das vacas, desempenho em produção de leite e eventos das lactações anteriores e da parição atual. Além disso, os dados foram coleatdos através de múltiplos sensores adaptados ao animal durante 3 sessões diárias para avaliar aspectos do leite, tais como porcentagem de gordura, de proteína, dados de lactose e conductividade, além de dados sobre total de passos dados, tempo de ruminação e tempo de repouso do animal.

Os dados considerados neste estudo devem ser usados para treinar, validar e testar modelos do Sistema de Monitoramento de Saúde Animal (AHMS) do projeto USDA da Dairy Laboratório de Biologia e Gerenciamento de Gado da Universidade de Cornell. Os dados contém informações sobre vacas e leite coletado. O conjunto de dados principal corresponde ao exame clínico diário dos animais são necessários

para analisar o período de uma semana antes até três semanas depois do parto em todos os animais. Diariamente, informações de uma vaca sobre um distúrbio de saúde são coletadas.

O estudo de Cornell compreende três fases, Fase I, Fase II e Fase III. O objetivo do estudo da Fase I é caracterizar o padrão de doença, avaliando como os parâmetros mudam com relação ao status do animal. No estudo da Fase II, o foco é desenvolver alertas que sirvam de indicadores aos agricultores quando uma vaca está doente, visando criar uma combinação de vários parâmetros que informam sobre a saúde do animal. A Fase III do estudo compreende a validação dos métodos usados com dados da Fase I, usando ferramentas desenvolvidas em campo em tempo real e, em seguida, avaliando falsos positivos e sensibilidade do método.

A descrição das variáveis do conjunto de dados se encontram na Tabela 1 e 2.

Objetivos

Neste estudo, os objetivos estão concentrados na Fase I, que busca caracterizar o padrão de doença, avaliando como os parâmetros se comportam com relação ao status do animal. As variáveis utilizadas para modelar a probabilidade do animal estar doente serão consideradas na forma escalar e de função e a ideia é identificar através do ajuste de modelos uma combinação de vários parâmetros que informem sobre a saúde do animal dentro de um determinado período de tempo. Por exemplo, é de interesse predizer a probabilidade de uma animal ficar doente no período de cinco dias após o parto a partir da curva de lactação do período. Para modelar a probabilidade de um animal ficar doente consideramos um modelo de regressão logística com covariáveis escalares (número de inseminações, histórico de doenças, etc) e funções de covariáveis, como por exemplo uma função dos dias em lactação do animal.

Tabela 1: Descrição das variáveis do conjunto de dados

Variável	Unidade	Descrição
ID	Número	Identificação da vaca
DiasLac	Dias	Dias em lactação (dia 0: dia do parto)
DiasRegimeFechado	Dias	Dias em regime Fechado
DiasSecos	Dias	Dias sem produzir leite
NumBezerro	Número	Número de bezerros nascidos no parto
ProdLeite	Gramas	Produção de leite
Gordura	%	Gordura no leite
Proteina	%	Proteína no leite
Lactose	%	Lactose no leite
Sangue	%	Sangue no leite
NumCelSomaticas	*1000/ml	Número de células somáticas no leite
TempoRepouso	Minutos	Tempo total de repouso por dia

Tabela 2: Descrição das variáveis do conjunto de dados - continuação

Variável	Unidade	Descrição
NumRepousos	Número	Número de repousos por dia
DuracaoMediaRepouso	Minutos	Duração média do cada repouso
Atividade	Passos/Hora	Total de passos dados no dia dividido por 24
RuminacaoUltimas24h	Minutos	Ruminação total nas últimas 24 horas
AlimentacaoUltimas24h	Gramas	Quantidade de comida ingerida nas últimas 24 horas
Hist_DistDigestivo	0-1	Há histórico de distúrbios digestivos - sim(1) não(0)
Hist_Mastite	0-1	Há histórico de mastite - sim(1) não(0)
Hist_Claudicao	0-1	Há histórico de claudicação - sim(1) não(0)
DiasNaoGravida	Dias	Dias não grávida
NInseminada	Número	Número de vezes que a vaca foi inseminada
DuracaoGestacao	Dias	Duração da gestação
DiasPrimeiraParicao	Dias	Idade quando pariu o primeiro bezerro
DuracaoLacAnterior	Dias	Duração da lactação anterior

1 Métodos Estatísticos

Seja Y_i o estado de saúde da vaca i (1 para doente e 0, caso contrário) em um determinado período, e seja $E(Y_i) = p_i$ a probabilidade do animal i estar doente nesse período. Inicialmente ajustamos o modelo de regressão logística (Agresti, 2003) da forma

$$\text{logito}(p_i) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij}, \tag{1}$$

onde x_{ij} representa a j-ésima variável explicativa da vaca i e β_j, o efeito desta variável no logito. Para essa abordagem, cada vaca possui uma única observação.

Posteriormente, ajustamos o modelo que acomoda as medidas repetidas de cada animal, representadas pelos dias entre 30 dias antes do parto e 40 dias após este evento. Consideramos um efeito aleatório para acomodar a variação intra-animal, devido a repetição de medições das vacas em tempos diferentes. Suponha Y_{ik} seja o estado de saúde da vaca i (1 para doente e 0, caso contrário) em um determinado dia k, e seja $E(Y_{ik}) = p_{ik}$ a probabilidade do animal i estar doente no dia k. Consideramos o ajuste de

$$\text{logito}(p_{ik}) = \beta_0 + \sum_{j=1}^{J} f_j(x_{ijk}) + u_i. \tag{2}$$

Aqui, $f_j(x)$ são funções suaves de $x_{ijk}\}$, k e j são referentes a observação do dia e a variável explicativa, respectivamente. A função $f_j(x)$ será estimada através de

$$\hat{f}_j(x) = \sum_{p=1}^{P} \hat{B}_p b_p(x), \tag{3}$$

onde b_p são as funções bases e $\hat{B}_p$ seus respectivos coeficientes; e $u_i \sim N(0, \sigma^2)$ é referente aos efeitos aleatórios para os animais analisados.

Para análise exploratória e avaliação da capacidade preditiva dos modelos, foi realizada uma separação aleatória dos animais em estudo em duas partes: 80% dos dados foram utilizados para o análise exploratória e ajuste e 20% para o teste de qualidade de predição do modelo. A forma de checagem da performance do modelo se deu através da curva ROC (Bradley, 1997) conjuntamente com a área sob a curva da mesma (Ballings and Van den Poel (2013)). A área sob a curva da curva ROC quantifica a capacidade do modelo em discriminar entre aqueles animais classificados como doentes dado que de fato são doentes e aqueles classificados como sadios quando na verdade foram observados como tal. Um modelo com capacidade de predição tão informativo quanto um classificador aleatório tem uma área igual a 0,5, já um modelo perfeito tem uma área igual a 1.

Neste trabalho, consideramos um nível de significância igual 0,05. Para realizar o ajuste da regressão logística em (1) utilizamos a própria função base *glm* do software R (R Core Team ,2019). O modelo em (2) foi ajustado por meio da função *gamm* do pacote *mgcv* do programa R. As funções suaves $f_j(x)$ foram estimadas através da expansão de bases de splines de regressão cúbica pela própria função *gamm*. Na função *gamm*, o modelo é ajustado através maximização da quasi-verossimilhança penalizada. Para mais informações sobre o ajuste com variáveis funcionais e efeitos aleatórios veja Wood (2017).

2 Estrutura e manipulação dos dados

O conjunto de dados utilizado pra este trabalho apresenta 71 observações para cada uma das 500 vacas estudadas, coletadas 30 dias antes do parto até 40 dias depois, resultando em um total de 35500 observações.

Além disso, o conjunto de dados apresentou diversos dados faltantes. Para contornar esse problema tomou-se o seguinte procedimento: se o animal apresentou poucos dados faltantes para uma variável, esses seriam substituidos pela média daquela variável ao longo das medidas repetidas do animal. Para o animal sem nenhuma informação numa determinada variável, o valor imputado foi a mediana da variável para as vacas com informação

3 Resultados

Foram utilizadas duas abordagens para resolver o problema de classificação. A primeira faz ajuste de um modelo a partir de uma única observação do estado de saúde do animal, como descrito na Seção 3.1. A segunda faz o ajuste de um modelo considerando medidas repetidas de cada animal, como descrito na Seção 3.2.

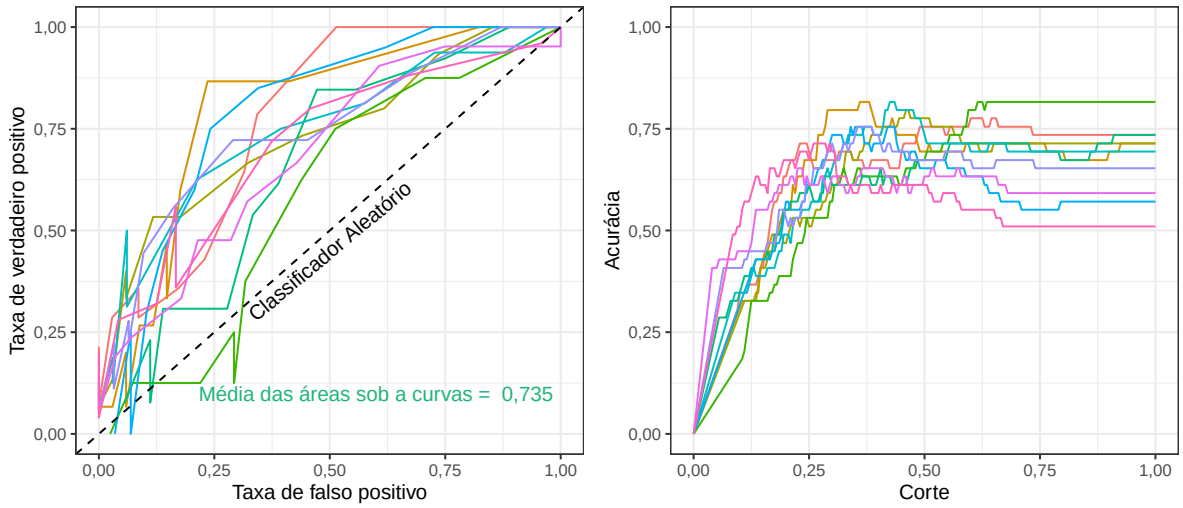

Figura 1: Curvas ROC e acurácias do modelo sem medidas repetidas alcançado por stepwise obtidas por meio de validação cruzada.

Tabela 3: Resultados do ajuste do modelo sem medidas repetidas obtido por stepwise.

Termo	Estimativa	Erro padrão	Estatística do teste	P-valor
Intercepto	15,344	5,963	2,573	<0,05
DiasSecos.media	-0,044	0,021	-2,098	<0,05
Gordura.media	0,606	0,201	3,013	<0,05
Proteina.media	-0,823	0,284	-2,893	<0,05
TempoRepouso.media	-0,002	0,001	-2,125	<0,05
NumRespousos.media	0,120	0,043	2,802	<0,05
RuminacaoUltimas24h.media	-0,009	0,003	-3,586	<0,05
AlimentacaoUltimas24h.media	-0,007	0,002	-3,653	<0,05
DiasNaoGravida.media	0,052	0,021	2,431	<0,05
DuracaoLacAnterior.media	-0,047	0,021	-2,219	<0,05

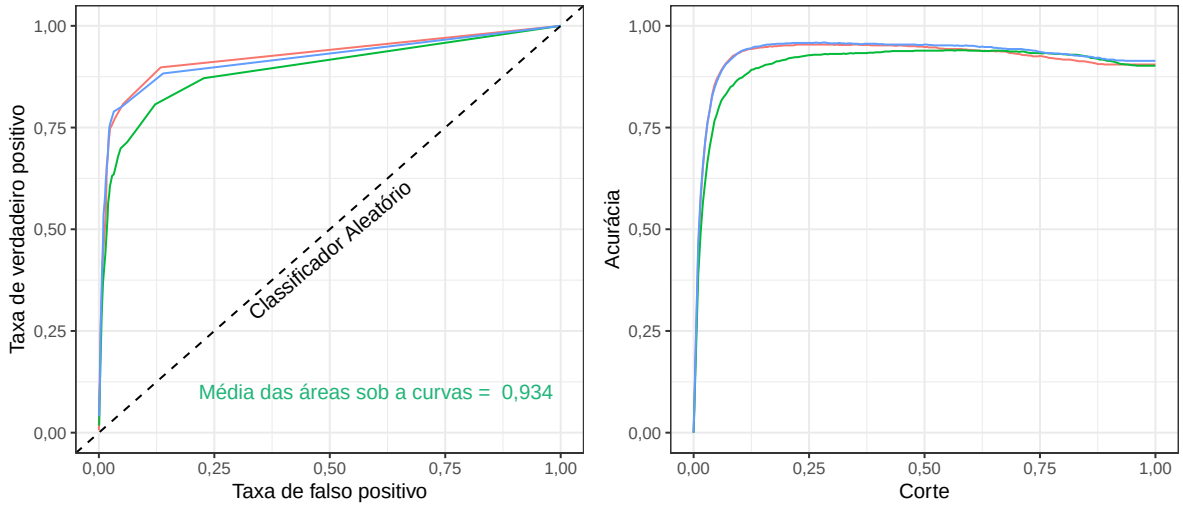

Figura 2: Curvas ROC e acurácias do modelo misto escalar obtidas por meio de validação cruzada.

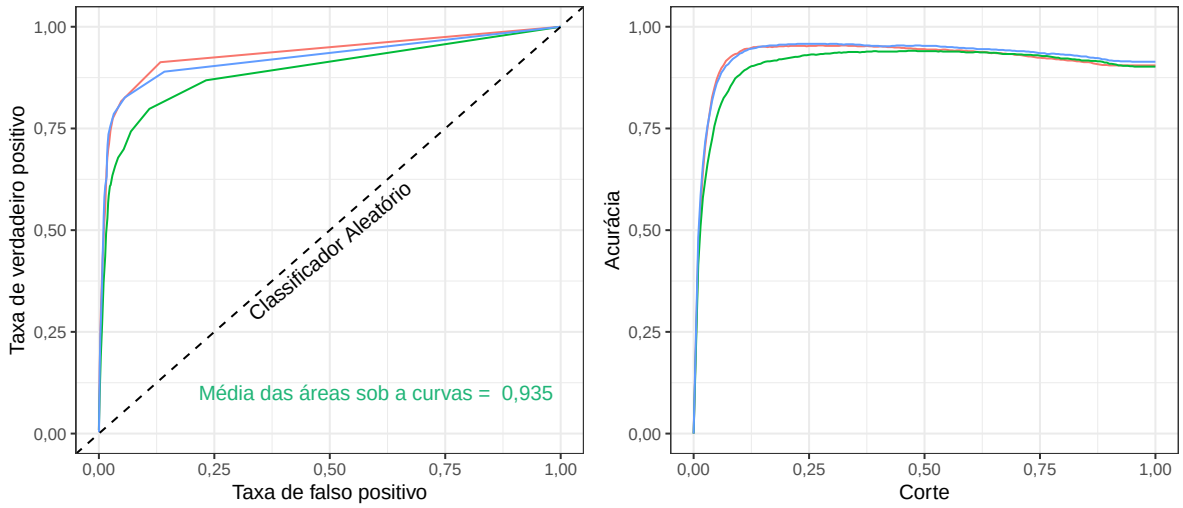

Figura 3: Curvas ROC e acurácias do modelo misto escalar-funcional obtidas por meio de validação cruzada.

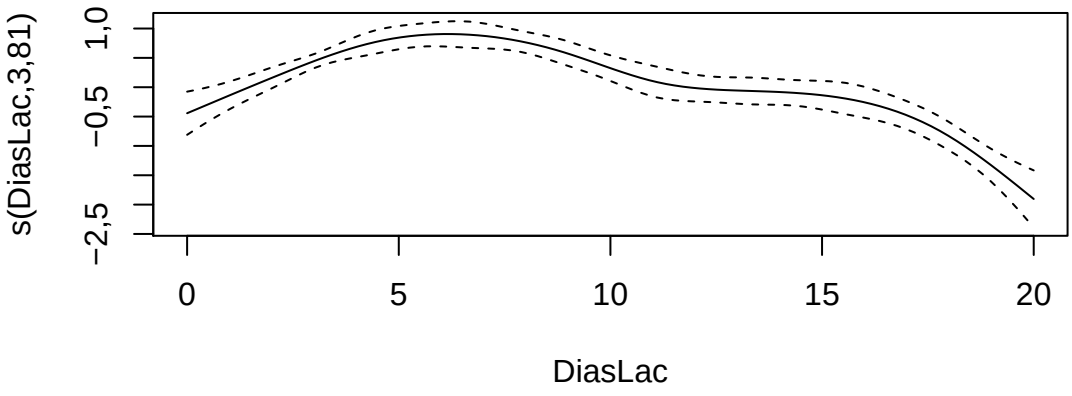

Figura 4: Curva estimada de dia de lactação (DiasLac)

Tabela 4: Estimativas dos parâmetros do ajuste do modelo misto escalar-funcional.

Termo	Estimativa	Erro padrão	Estatística do teste	P-valor
Intercepto	<0,001	0,513	-6,516	<0,001
RuminacaoUltimas24h	0,003	0,002	1,731	0,084
RuminacaoUltimas24h_ontem	0,005	0,002	2,957	0,003
AlimentacaoUltimas24h	<0,001	0,002	-0,330	0,741
AlimentacaoUltimas24h_ontem	0,004	0,002	1,802	0,072
doente_hoje1	2,692	0,148	18,144	<0,001
doente_ontem1	0,913	0,163	5,618	<0,001
ProdLeite	<0,001	<0,001	1,166	0,243
ProdLeite_ontem	<0,001	<0,001	1,370	0,171
RuminacaoUltimas24h:RuminacaoUltimas24h_ontem	<0,001	<0,001	-4,592	<0,001
AlimentacaoUltimas24h:AlimentacaoUltimas24h_ontem	<0,001	<0,001	-2,822	0,005
doente_hoje1:doente_ontem1	0,478	0,217	2,204	0,028
ProdLeite:ProdLeite_ontem	<0,001	<0,001	-1,808	0,071

Tabela 5: P-valor do efeito da função de suavização s(DiasLac) no modelo misto escalar-funcional. GLE é referente aos graus de liberdade efetivos.

Termo	GLE	Estatística do teste	P-valor
s(DiasLac)	3,811	36,197	<0,001

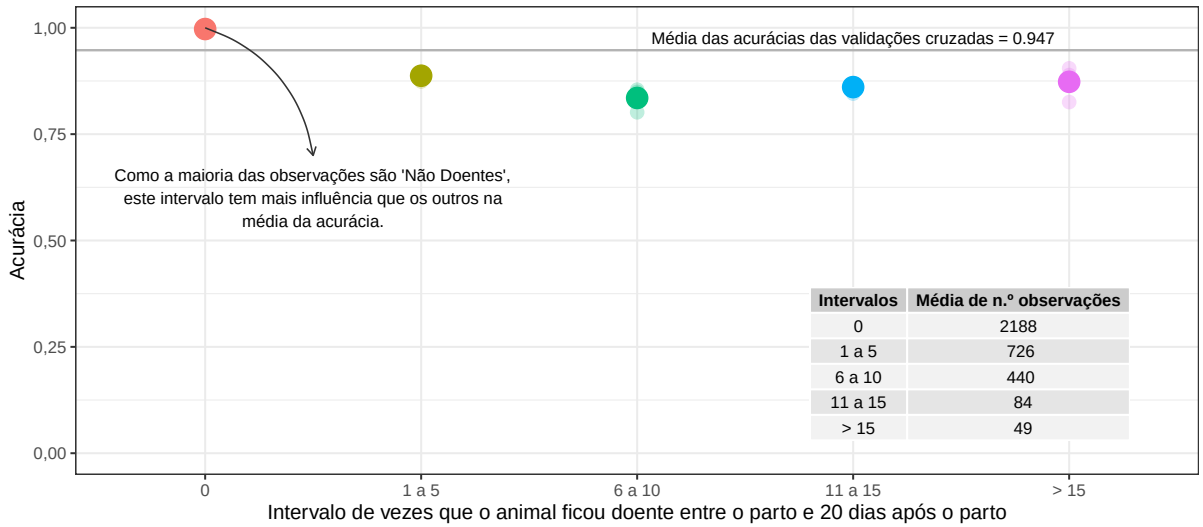

Figura 5: Acurácia em intervalos de quantas vezes as vacas do conjunto de teste ficaram doentes entre o dia do parto e os 20 dias posteriores, obtida pela validação cruzada.

Agresti, Alan. 2003. *Categorical Data Analysis*. Vol. 482. John Wiley & Sons.

Ballings, Michel, and Dirk Van den Poel. 2013. *AUC: Threshold Independent Performance Measures for Probabilistic Classifiers.* https://CRAN.R-project.org/package=AUC.

Bozdogan, H. 1987. “Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions.” *Psychometrika* 52: 345–70.

Bradley, Andrew P. 1997. “The Use of the Area Under the Roc Curve in the Evaluation of Machine Learning Algorithms.” *Pattern Recognition* 30 (7): 1145–59.

Drackley, James K. 1999. “Biology of Dairy Cows During the Transition Period: The Final Frontier?” *Journal of Dairy Science* 82 (11): 2259–73. https://doi.org/https://doi.org/10.3168/jds.S0022-0302(99)75474-3.

Goff, JP, and RL Horst. 1997. “Physiological Changes at Parturition and Their Relationship to Metabolic Disorders1, 2.” *Journal of Dairy Science* 80 (7): 1260–8.

Kaufman, EI, VH Asselstine, SJ LeBlanc, TF Duffield, and TJ DeVries. 2018. “Association of Rumination Time and Health Status with Milk Yield and Composition in Early-Lactation Dairy Cows.” *Journal of Dairy Science* 101 (1): 462–71.

Lago, Alexandre Vaz AND Susin, Ernani Paulino do AND Pires. 2001. “Efeito da condiÃ§Ãcorporal ao parto sobre alguns parÃdo metabolismo energÃ, produÃ§Ãde leite e incidÃde doenÃno pÃ-parto de vacas leiteiras.” *Revista Brasileira de Zootecnia* 30 (October): 1544–9. http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-35982001000600023&nrm=iso.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wilcoxon, Frank. 1945. “Individual Comparisons by Ranking Methods.” *Biometrics Bulletin* 1 (6): 80–83. http://www.jstor.org/stable/3001968.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. CRC press.

Yamashita, Toshie, Keizo Yamashita, and Ryotaro Kamimura. 2007. “A Stepwise Aic Method for Variable Selection in Linear Regression.” *Communications in Statistics - Theory and Methods* 36 (13): 2395–2403. https://doi.org/10.1080/03610920701215639.