Pecuária leiteira: Avaliando o efeito de um aditivo na média diária de gordura no leite produzido

Pedro Lima Garcia, junho de 2023

Introdução

A pecuária de leite exerce um papel muito importante para a economia mundial. Mais de 600 milhões de pessoas em todo o mundo vivem em cerca de 133 milhões de fazendas leiteiras, a maioria pequenas propriedades, que abrigam em média de duas a três vacas. Mesmo realizado principalmente por pequenos produtores, a atividade representa 40% do valor do rendimento agrícola mundial. (LATICÍNIOS HOLANDÊS, 2021)

Só no Brasil, 415.000 pessoas vivem da pecuária, das quais 250.000 estão no mercado formal (que passa pela inspeção sanitária). Desse subtotal, quase 170.000 são considerados pequenos produtores (que ordenham até 250 litros ao dia), de acordo com a Associação Brasileira dos Produtores de Leite e a consultora MilkPoint. Por isso, essa atividade tem um valor inestimável para a subsistência de muitas famílias.

Isso posto, é de suma importância que técnicas de modelagem estatística sejam utilizadas para analisar os dados e melhorar a produtividade. "Quando o produtor tem pouco conhecimento, o leite é tratado como algo simples e o animal acaba dando só cinco ou seis litros por dia, em vez dos 20 que poderia oferecer", diz o agrônomo Sergio Diehl.

O presente trabalho busca avaliar o efeito de um aditivo na média diária de gordura no leite produzido. Para isso, utilizou-se dados de um experimento agrícola que foi conduzido em 50 vacas. Quatro dietas (tratamentos) foram consideradas, correspondendo a diferentes níveis do aditivo, e três variáveis foram registradas antes da atribuição do tratamento: número de lactação, idade e peso inicial da vaca. Após o tratamento, mediu-se o peso final das vacas, a média diária de ração consumida e as quantidades de sólidos, proteína e gordura no leite.

Os dados e sua documentação estão disponíveis no repositório de datasets dos livros Data Analysis Using Regression and Multilevel/Hierarchical Models e Regression and Other Stories, ambos escritos pelos autores Andrew Gelman e Jennifer Hill.

Metodologia

O trabalho foi todo realizado em linguagem R no software RStudio, incluindo a implementação dos modelos, a produção dos plots e os cálculos em geral. Fez-se uso dos pacotes GGally, rstanarm, ggplot2, tidyverse, lme4, caret e psych. Os modelos implementados foram lineares com múltiplos preditores, com e sem interações entre as variáveis.

Baseado na documentação dos dados, supôs-se que as variáveis que poderiam ter influência sobre a média diária de gordura no leite produzido (além do aditivo) seriam a idade da vaca, o peso da vaca e a média diária de ração consumida. Vale ressaltar que o efeito do aditivo poderia ser incluído ao modelo de (pelo menos) duas maneiras distintas: pela porcentagem de aditivo na dieta da vaca (level) ou pela média diária de aditivo consumido (dry*level). O primeiro passo foi decidir entre as duas abordagens. Para isso, foram ajustados os seguintes modelos:

```
> print(M1)

Call:
lm(formula = mean ~ level, data = vacas)

Coefficients:
(Intercept)        level
      1.943        1.251

> print(M2)

Call:
lm(formula = mean ~ aditivo, data = vacas)

Coefficients:
(Intercept)      aditivo
     1.8716       0.1046
```

Sabe-se que os coeficientes estão atrelados às unidades de medida das variáveis. Para poder compará-los em magnitude, fez-se uma padronização com a função lm.beta([modelo]) na qual ficou claro que a variável aditivo tem mais influência sobre a média de gordura do que a variável level (o que faz sentido intuitivamente).

```
> lm.beta(M1)
    level
0.3062717
> lm.beta(M2)
  aditivo
0.4504245
```

A variância explicada por cada modelo também indica o aditivo como uma variável preferível, já que em seu modelo $R^2 \approx 0.186$, enquanto no modelo do level $R^2 \approx 0.074$ (quase nada). Portanto, a variável aditivo foi a escolhida.

Os ajustes consistiram em partir desse modelo mais simples (com um único preditor) e adicionar preditores de acordo com o aumento da quantidade de variância explicada pelo modelo (preditores que apresentaram influência muito baixa foram desconsiderados). Uma vez selecionadas todas as variáveis relevantes, passou-se a adicionar ao modelo possíveis interações entre elas (como idade e peso, peso e ração consumida, entre outros).

Para comparar os modelos, utilizou-se os métodos AIC e BIC, ambos com funções nativas do R. Desse modo, o melhor modelo pôde estimar os coeficientes dos preditores considerados relevantes e consequentemente avaliar suas influências sobre a média diária de gordura no leite. Calculou-se também os intervalos de confiança para cada um desses coeficientes, com a função confint([modelo]).

Análise Exploratória de Dados

Os dados originais consistem numa tabela de 50 linhas (uma para cada vaca) e 10 colunas: level (quantidade de aditivo na ração caracterizando o tratamento), lactation (número de vezes que a vaca lactou), age (idade em semanas), initial.weight (peso em libras no início do tratamento), dry (média diária de ração consumida), milk (média diária de leite produzido), fat (porcentagem de gordura no leite), solids (porcentagem de sólidos no leite, isto é, gordura, proteína, lactose e minerais), final.weight (peso em libras no final do tratamento) e protein (porcentagem de proteína no leite).

O primeiro passo foi criar a coluna "mean" com o produto milk*fat/100, que corresponde à média diária de gordura no leite produzido, quantidade sobre a qual se desejava avaliar o efeito do aditivo. Além disso, tornou-se a coluna level categórica, adicionando uma coluna "diet" que associa as quantidades de aditivo 0.0, 0.1, 0.2 e 0.3 às dietas 0, 1, 2 e 3, respectivamente. Como citado na seção anterior, criou-se também a coluna "aditivo" com o produto level*dry. O resultado foi a tabela cujo trecho segue abaixo na Figura 1:

	lactation	age	initial.weight	dry	milk	fat	solids	final.weight	protein	mean	level	diet	aditivo
1	3	49	1360	15.429	45.552	3.88	8.96	1442	3.67	1.767418	0.0	diet 0	0.0000
2	3	47	1498	18.799	66.221	3.40	8.44	1565	3.03	2.251514	0.0	diet 0	0.0000
3	2	36	1265	17.948	63.032	3.44	8.70	1315	3.40	2.168301	0.0	diet 0	0.0000
4	2	33	1190	18.267	68.421	3.42	8.30	1285	3.37	2.339998	0.0	diet 0	0.0000
5	2	31	1145	17.253	59.671	3.01	9.04	1182	3.61	1.796097	0.0	diet 0	0.0000

Figura 1: 5 primeiras linhas da tabela "vacas".

Com os dados tratados foram feitos então alguns plots preliminares, afim de observar algum indicativo de padrão ou tendência geral.

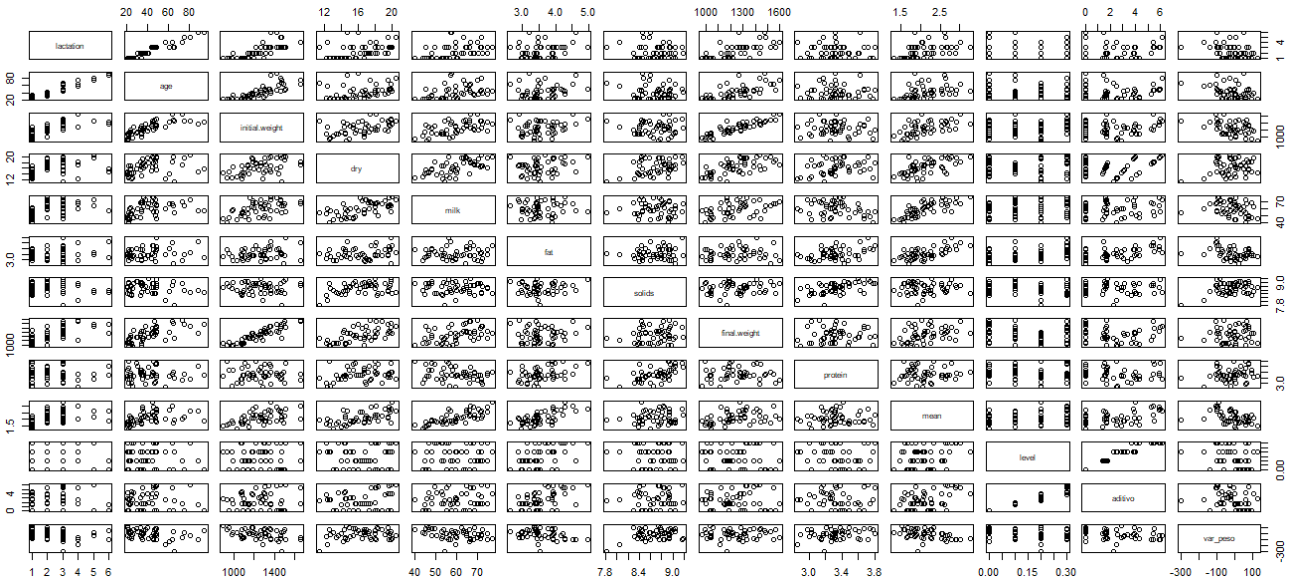

Figura 2: Plot pairs(vacas).

Os que pareceram mais interessantes a princípio foram o boxplot “diet x mean” e o scatterplot “level x aditivo”.

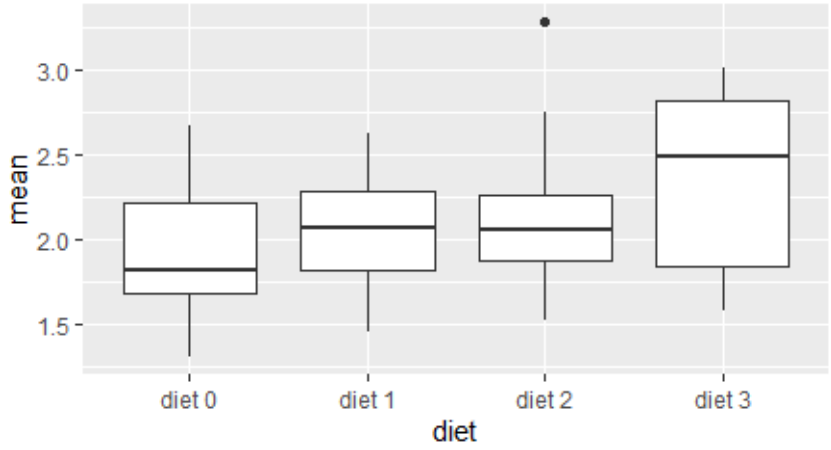

Figura 3: Boxplot “diet x mean”.

No boxplot da Figura 2 notou-se que, embora as médias de gordura sejam mais dispersas na dieta 3, sua mediana é bem superior em relação às demais, sugerindo que o aditivo aumenta a média de gordura no leite.

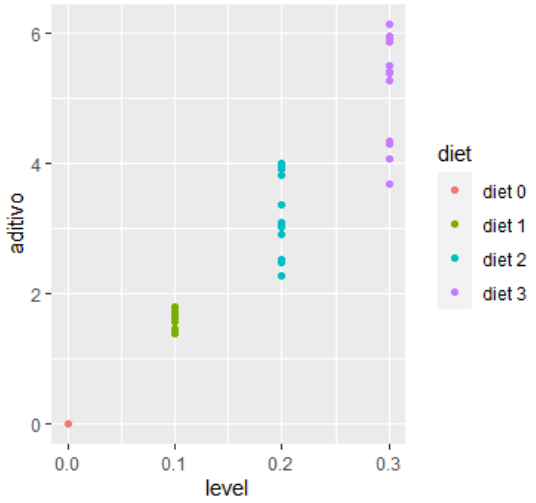

Figura 4: scatter plot “level x aditivo”.

No scatterplot da Figura 3 verificou-se que, com exceção de uma única vaca, vacas submetidas à dietas com maior porcentagem de aditivo consumiram mais aditivo que vacas submetidas à dietas com menor porcentagem de aditivo. Isso poderia não ocorrer se houvessem muitas vacas com alta porcentagem de aditivo em sua dieta que consumissem pouquíssima ração. Como não foi o caso, a variável level pôde ser descartada, pois seu efeito já seria explicado pela variável aditivo.

Resultados

Primeiramente foram calculados os valores de AIC e o BIC dos seguintes modelos:

```
> M3 <- lm(mean ~ initial.weight + age + dry + aditivo, vacas)
> M4 <- lm(mean ~ initial.weight + age + aditivo, vacas)
> M5 <- lm(mean ~ initial.weight + dry + aditivo, vacas)
> M6 <- lm(mean ~ initial.weight + aditivo, vacas)
> M7 <- lm(mean ~ age + dry + aditivo, vacas)
> M8 <- lm(mean ~ age + aditivo, vacas)
> M9 <- lm(mean ~ dry + aditivo, vacas)
> AIC(M3, M4, M5, M6, M7, M8, M9)
   df      AIC
M3  6 36.85059
M4  5 43.62589
M5  5 34.92406
M6  4 41.63242
M7  5 37.63937
M8  4 48.49594
M9  4 40.66784
> BIC(M3, M4, M5, M6, M7, M8, M9)
   df      BIC
M3  6 48.32272
M4  5 53.18601
M5  5 44.48418
M6  4 49.28051
M7  5 47.19949
M8  4 56.14403
M9  4 48.31593
```

Ambos os critérios apontaram o modelo M5 (que considera o peso inicial, a média diária de ração consumida e a quantidade de aditivo consumido por dia) como o melhor. Surpreendentemente, incluir a variável age (idade) piora o modelo, pois não só aumenta os valores de AIC e BIC como também diminui o R^2. Isso talvez se deva ao fato de que a diferença de idade entre as vacas observadas não chega a 1 ano e meio.

Escolhidas as variáveis mais relevantes, tentou-se então melhorar o modelo adicionando possíveis interações entre elas e calculando novamente os valores de AIC e BIC.

```
> M10 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:dry + initial.weight:aditivo + dry:aditivo, vacas)
> M11 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:dry + initial.weight:aditivo, vacas)
> M12 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:dry + dry:aditivo, vacas)
> M13 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:aditivo + dry:aditivo, vacas)
> M14 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:dry, vacas)
> M15 <- lm(mean ~ initial.weight + dry + aditivo + initial.weight:aditivo, vacas)
> M16 <- lm(mean ~ initial.weight + dry + aditivo + dry:aditivo, vacas)
> AIC(M10, M11, M12, M13, M14, M15, M16)
    df      AIC
M10  8 37.17157
M11  7 36.70779
M12  7 35.18340
M13  7 36.70861
M14  6 35.38819
M15  6 36.57636
M16  6 34.84167
> BIC(M10, M11, M12, M13, M14, M15, M16)
    df      BIC
M10  8 52.46775
M11  7 50.09195
M12  7 48.56757
M13  7 50.09277
M14  6 46.86033
M15  6 48.04849
M16  6 46.31381
```

O modelo que apresentou melhores resultados foi então o M16 (que considera o peso inicial, a média diária de ração consumida, a quantidade de aditivo consumido por dia e a interação entre ração e aditivo). Porém, ao compará-lo com o M5, houve pouca diferença entre seus valores de AIC e BIC (ocorrendo até uma sutil piora em relação ao BIC). Também houve pouco aumento no R^2, como mostra a Tabela 1. Logo, optou-se pelo modelo mais simples, o M5.

Modelo	AIC	BIC	R^2
M5	34.92	44.48	0.49
M16	34.84	46.31	0.50

Tabela 1: AIC, BIC e R^2 dos modelos M5 e M16.

Por fim, foram calculados intervalos de confiança de 95% para os coeficientes do modelo M5. Nota-se que nenhum dos intervalos contém o 0, o que faz sentido com a interpretação dos coeficientes.

```
> confint(M5)
                    2.5 %      97.5 %
(Intercept)    -0.9054592255 0.595911602
initial.weight  0.0002215256 0.001390357
dry             0.0212252651 0.111605553
aditivo         0.0232528210 0.122493564
```

Conclusão

Considerando a complexidade e a capacidade do modelo em explicar a variância presente nos dados, o modelo que se mostrou preferível foi o M5, que considerou o peso inicial das vacas, a média diária de ração consumida e a média diária de aditivo consumido. O modelo foi capaz de explicar aproximadamente 49% da variância dos dados e mostrou que o aditivo tem um efeito positivo na média diária de gordura no leite. Tal efeito pode ser interpretado nos coeficientes da seguinte maneira: mantidos constantes o peso da vaca e a média diária de ração consumida, 1ml a mais de aditivo aumenta a média diária de gordura no leite em 7ml. Afirmar se esse efeito é significativo ou não exige conhecimento específico da área e foge ao escopo desse trabalho.

A principal limitação do trabalho foi a quantidade reduzida de dados, com apenas 50 observações. Além de poucas, as dietas ainda as subdividiam em grupos de 12 ou 13 observações, motivo pelo qual jugou-se não fazer sentido avaliar o efeito do aditivo de modo hierárquico, mas sim de modo geral, transformando a variável level na variável aditivo (como explicado na seção de Metodologia). Também pela escassez de dados, o melhor modelo conseguiu explicar apenas metade da variância presente nos dados, o que, embora não tenha sido determinante para avaliar a qualidade do modelo, foi um pouco decepcionante. De todo modo, como dito na seção de introdução, fazendas não costumam ter muitas vacas leiteiras na vida real, o que torna a situação de poucos dados um problema válido e aplicável.

Do ponto de vista técnico, o trabalho contribuiu consideravelmente no aprendizado da linguagem R como ferramenta não só de visualização (como já havia sido vista em disciplinas anteriores do curso), mas principalmente de modelagem estatística, com a qual foi possível aplicar a teoria passada em aula. De maneira particular, foi interessante ver na prática que é possível que o R^2 diminua com a adição de uma variável ao modelo (houve uma discussão na aula sobre como isso seria possível, já que é uma ideia contraintuitiva - e acabei vendo acontecer nesse trabalho), como citado na seção Resultados.

Um direcionamento futuro possível seria buscar entender as colunas da tabela que não estavam explicadas na documentação do dataset (como solids, por exemplo) por pesquisa própria. Dessa maneira, o modelo poderia ser ajustado considerando a adição de novas variáveis e de interações entre elas. Além disso, o efeito do aditivo poderia ser avaliado em outras variáveis além da gordura no leite, como na variação do peso das vacas ao longo do tratamento ou na média diária de proteína no leite.

Referências

[1] GELMAN, Andrew.; HILL, Jennifer. **Data Analysis Using Regression and Multilevel/Hierarchical Models.** 1st edition. Cambridge: Cambridge University Press, 2006.

[2] GELMAN, Andrew.; HILL, Jennifer.; VEHTARI, Aki. **Regression and Other Stories.** 2nd edition. Cambridge: Cambridge University Press, 2022.

[3] A importância do leite para a economia mundial. **Laticínios Holandês**, 2021. Disponível em: <https://laticiniosholandes.com.br/a-importancia-do-leite-para-a-economia-mundial/>. Acesso em: 17 de junho de 2023.

[4] O iogurte que mudou a vida de uma comunidade rural. **The World Bank**, 2014. Disponível em: <https://www.worldbank.org/pt/news/feature/2014/01/03/brazil-sao-paulo-milk-yoghurt-production-changes-lives-rural-community>. Acesso em: 27 de junho de 2023.