Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Uso de Modelos de Regressão Não-Linear em Dados de Leite Materno

Gabriella Luz Firmino Mourão

Ouro Preto-MG
Julho – 2021

Gabriella Luz Firmino Mourão

Uso de Modelos de Regressão Não-Linear em Dados de Leite Materno

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador(a)

Prof. Dr. Eduardo Bearzoti

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto-MG

Julho – 2021

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Gabriella Luz Firmino Mourão

Uso de Modelos de Regressão Não-Linear em Dados de Leite Materno

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 16 de julho de 2021

Membros da banca

Dr. Eduardo Bearzoti - Orientador - Universidade Federal de Ouro Preto
Dra. Carolina Silva Pena - Universidade Federal de Ouro Preto
Dr. Marcelo Carlos Ribeiro - Universidade Federal de Ouro Preto

Professor Dr. Eduardo Bearzoti, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 16/07/2021

Documento assinado eletronicamente por **Eduardo Bearzoti**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 16/07/2021, às 15:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Carolina Silva Pena**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 20/07/2021, às 10:56, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Carlos Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 21/07/2021, às 15:44, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0194564** e o código CRC **B5C679F4**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

Agradeço a Deus, aos meus pais, por todo apoio e amor, e ao meu orientador Eduardo.
Este trabalho é inteiramente dedicado a eles. Gratidão.

Agradecimentos

Em primeiro lugar, agradeço a Deus, por ter me ajudado e me abençoado em todo este trajeto. Obrigada, Deus, por me mostrar que eu tinha força, até nos momentos mais difíceis. Ao meu pai, por toda a dedicação, e por sempre me mostrar o caminho certo; eu não seria nada sem você, pai. À minha tia Mirian, pelo amor incondicional e devoção sem igual. À minha mãe, por todo o amor do mundo. Às minhas irmãs, Juliana e Jéssica, por serem minhas companheiras de toda uma vida. À minha irmã Isabella, por trazer juventude para minha vida novamente. Agradeço também aos meus avós por terem me proporcionado tanto amor e carinho. A todos os meus tios que me apoiaram e me ajudaram ao longo dessa caminhada. Ao meu cunhado Ygor, por ser companheiro até nos piores momentos. Essa vitória é nossa.

Ao meu orientador Eduardo Bearzoti, por toda ajuda e dedicação a esse trabalho; sem você esse trabalho não seria possível. À professora Camila Menezes, por gentilmente disponibilizar os dados utilizados neste trabalho. A todos os professores do DEEST, por sempre serem solícitos, e à UFOP, pelo ensino de qualidade e gratuito.

Aos meus amigos de curso, Leilayne, Pedro, Willian, Denis e Marcos, por todo o companheirismo nesta jornada. Vocês trouxeram leveza e risos para os meus dias. Aos meus amigos de vida, e aos que tive a oportunidade de conhecer ao longo desta jornada, por compartilharem os melhores anos de faculdade comigo.

Por fim, agradeço aos meus companheiros de vida e fiéis companheiros caninos, Tchutchu, Ringo, Lua e Bethoven. Vocês trouxerem luz e alegria para minha vida em todos os momentos.

"O pensamento cresce, cresce e toma conta de toda a nossa cabeça e nosso coração.
Vive em nossos olhos e em tudo que é pedaço da vida da gente."

José Mauro de Vasconcelos

Uso de Modelos de Regressão Não-Linear em Dados de Leite Materno

Autora: Gabriella Luz Firmino Mourão
Orientador: Prof. Dr. Eduardo Bearzoti

Resumo

O uso de modelos de regressão corresponde a uma das mais utilizadas dentre as metodologias estatísticas, sendo que em geral maior ênfase é dada aos modelos lineares e lineares generalizados. Contudo, a classe dos modelos de regressão não lineares corresponde a uma alternativa igualmente de grande utilidade, dada a maior flexibilidade de funções relacionando a variável preditora com a variável resposta. Os modelos não lineares são utilizados nas mais variadas áreas de conhecimento, como Econometria, cinética química, experimentos agrícolas, dentre outras. Este trabalho teve por objetivo descrever e ilustrar esta classe de modelos, aplicando-os em dados de leite humano, oriundos de uma linha de pesquisa da Escola de Nutrição da UFOP. Foram considerados dois modelos não-lineares, admitindo-se distribuição normal para os resíduos, e com pequenas modificações em relação aos modelos conforme originalmente propostos. Em um primeiro momento, foi utilizado um modelo de regressão logística normal para estimar o número de microrganismos presentes em alíquotas de leite, em função da temperatura de processamento do leite. As modificações aqui compreenderam a inclusão de um termo constante, bem como a possibilidade de a curva ter um formato de "S" invertido. A regressão logística apresentou grande aderência aos dados, e permitiu identificar um método de processamento potencialmente superior. Em um segundo momento, foi ajustado um modelo bi-segmentado contínuo, utilizando-se uma variável *dummy*, com ponto de interseção entre os segmentos desconhecido, o que torna o modelo não linear. Foram analisados dados referentes ao teor de uma fração lipídica do leite (o hexanal), em função do tempo de armazenamento de leite materno. Aqui foi feita outra pequena adaptação, admitindo-se a possibilidade de que a relação funcional entre a variável preditora e a variável resposta pudesse corresponder a outras funções que não apenas a função afim. Esta flexibilização melhorou sobremaneira a qualidade do ajustamento,

e permitiu descrever o comportamento do hexanal em função do tempo, evidenciando um segmento de plena ascensão, seguido de um outro segmento com decaimento progressivo. Estes resultados ilustram como, em uma mesma linha de pesquisa, diferentes modelos de regressão não linear foram de grande utilidade, evidenciando a flexibilidade e o potencial desta classe de modelos de regressão.

Palavras-chave: regressão não linear, regressão logística, modelos segmentados, leite materno.

Use of Nonlinear Regression Models in Breast Milk Data

Author: Gabriella Luz Firmino Mourão
Advisor: Prof. Dr. Eduardo Bearzoti

Abstract

The use of regression models corresponds to one of the most used among statistical methodologies, and in general greater emphasis is given to linear and generalized linear models. However, the class of nonlinear regression models corresponds to an alternative equally of great utility, given the greater flexibility of functions relating the predictor variable to the response variable. Nonlinear models are used in the most varied areas of knowledge, such as Econometrics, chemical kinetics, agricultural experiments, among others. This work aimed to describe and illustrate this class of models, applying them to human milk data, from a research program of the School of Nutrition of UFOP. Two non-linear models were considered, assuming normal distribution for the residuals, and with small modifications in relation to the models as originally proposed. At first, a normal logistic regression model was used to estimate the number of microorganisms present in milk aliquots, as a function of the milk processing temperature. The modifications here included a constant term in the model, as well as the possibility of the curve having an inverted S shape. Logistic regression showed great adherence to the data, and allowed the identification of a potentially superior processing method. In a second moment, a continuous bi-segmented model was fitted, using a dummy variable, with an unknown intersection point between the segments, which makes the model non-linear. Data related to the content of a lipid fraction in milk (the hexanal) were analyzed as a function of the storage time of breast milk. Here, another small adaptation was made, admitting the possibility that the functional relationship between the predictor variable and the response variable could correspond to functions other than just the affine function. This flexibility greatly improved the quality of the adjustment, and made it possible to describe the behavior of the hexanal as a function of time, showing a segment of full rise, followed by another segment with progressive decay. These results illustrate how, in a unique research program, different nonlinear

regression models were quite useful, showing the flexibility and potential of this class of regression models.

Keywords: nonlinear regression, logistic regression, segmented models, breast milk.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

Os chamados modelos de regressão são utilizados para descrever o comportamento de variáveis de interesse e a relação entre elas, sendo, provavelmente, uma das mais utilizadas dentre as metodologias estatísticas.

As categorias de modelos de regressão mais utilizadas poderiam simplificadamente ser classificadas em: lineares, não lineares e lineares generalizados. Em geral, quando se fala em modelos de regressão lineares e não lineares, admite-se uma distribuição normal para os resíduos (e consequentemente para a variável resposta). Os modelos lineares contêm, em sua parte estrutural, parâmetros que guardam uma relação de linearidade entre si (embora não necessariamente as variáveis explicativas apresentem tal relação), enquanto os modelos não lineares apresentam relações mais gerais entre seus parâmetros. Os modelos de regressão lineares generalizados admitem outras distribuições para a variável resposta, que não a normal, sendo seu valor esperado uma função de um preditor linear de parâmetros.

Embora os modelos de regressão lineares correspondam possivelmente à categoria mais utilizada, nem sempre propiciam ajustes satisfatórios, especialmente quando a relação existente entre as variáveis explicativas e a variável resposta apresenta uma natureza de maior complexidade. Nestas situações, os modelos de regressão não linear podem corresponder a uma alternativa conveniente, dada sua grande versatilidade, em virtude da gama muito maior de funções que permitem levar em conta. Os modelos não lineares são utilizados nas mais variadas áreas de conhecimento, como Econometria, cinética química, experimentos agrícolas, dentre outras.

Enquanto que os modelos de regressão lineares e lineares generalizados costumam ser amplamente abordados em cursos de graduação em Estatística, nem sempre é dada muita ênfase aos modelos de regressão não linear. Este trabalho teve assim como objetivo apresentar uma breve descrição desta classe de modelos, em especial aos procedimentos para o seu ajustamento, e ilustrá-la com dois modelos em particular: a regressão logística com distribuição normal para os resíduos, e os chamados modelos segmentados. Ambos

foram exemplificados utilizando dados reais referentes a características do leite materno.

2 Referencial Teórico

A análise de regressão é uma das técnicas mais utilizadas nas mais diversas áreas de estudo. Nesta técnica, o modelo mais simples é o da chamada regressão linear simples (ver, por exemplo, HOFFMANN, 2016), no qual há uma única variável independente, podendo ser representado por:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

no qual Y_i é o valor observado para a variável dependente, β_0 o intercepto, β_1 o coeficiente angular da reta a ser ajustada, X_i o valor observado da variável explicativa, e ϵ_i é o erro, ou resíduo, desconhecido, do modelo, admitido como tendo distribuição normal, com média 0 e variância constante σ^2.

Em algumas situações, há o interesse em se estudar o efeito de mais de uma variável explicativa sobre uma mesma variável resposta. Por exemplo, em um estudo sobre a gordura corporal de indivíduos, esta poderia ser considerada como função tanto da ingestão de comida industrializada como da frequência de práticas de exercício físico. Quando se tem mais de uma variável preditora, como numa situação como essa, tem-se os chamados modelos de regressão linear múltipla (HOFFMANN, 2016; DRAPER & SMITH, 2014). Por exemplo, com duas variáveis preditoras, este modelo poderia ser representado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{2.2}$$

Ambos os modelos 2.1 e 2.2 são *lineares*. Um modelo é dito linear (em relação aos seus parâmetros) quando todas as derivadas de Y_i em relação aos parâmetros não envolvem nenhum parâmetro. Por exemplo, o modelo 2.2 é linear, pois as derivadas parciais:

$$\frac{\partial Y_i}{\partial \beta_0} = 1 \qquad \frac{\partial Y_i}{\partial \beta_1} = X_{1i} \qquad \frac{\partial Y_i}{\partial \beta_2} = X_{2i}$$

não envolvem nenhum parâmetro.

Os modelos de regressão linear simples e múltipla são geralmente ajustados pelo método dos quadrados mínimos, que coincide com o método da máxima verossimilhança,

quando se admite distribuição normal para os resíduos. Um modelo de regressão múltipla contendo um intercepto e um certo número de variáveis preditoras pode ser expresso matricialmente através de:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.3}$$

em que $\mathbf{y}$ é o vetor contendo os n valores Y_i. A matriz $\mathbf{X}$ contem n linhas, cada qual referente a uma observação i; a primeira coluna contém valores iguais a 1 (referentes ao intercepto), e as demais colunas contêm os valores de cada uma das variáveis preditoras. O vetor $\boldsymbol{\beta}$ contém os parâmetros do modelo, e o vetor $\boldsymbol{\epsilon}$ é o vetor de resíduos. Pode-se demonstrar que o método dos quadrados mínimos consiste em obter a solução do chamado *sistema de equações normais*, dado por:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{2.4}$$

Quando a relação estrutural entre Y_i e as variáveis preditoras não satisfaz a condição de linearidade, tem-se um modelo de regressão não linear. Nesta classe de modelos, admite-se em geral que os resíduos sejam aditivos, e assim tais modelos, de forma geral, poderiam ser representados por:

$$Y_i = f(\boldsymbol{\xi}_i, \boldsymbol{\beta}) + \epsilon_i \tag{2.5}$$

no qual $\boldsymbol{\xi}_i$ representa um vetor de valores de um certo número de variáveis preditoras para a observação i, e $\boldsymbol{\beta}$ um vetor de parâmetros. Aqui se está considerando que pelo menos uma derivada parcial de Y_i em relação aos elementos de $\boldsymbol{\beta}$ dependa de algum elemento deste mesmo vetor.

Da mesma maneira que nos modelos de regressão linear, os métodos mais comuns para a estimação dos elementos de $\boldsymbol{\beta}$ são o método dos quadrados mínimos e o da máxima verossimilhança, sendo que estes métodos coincidem, quando admite-se que os resíduos tenham distribuição normal e variância constante. Os modelos não lineares, contudo, têm a particularidade de que tais métodos conduzem a sistemas de equações não lineares, sendo necessários métodos numéricos para sua resolução. A seguir será apresentado um dos métodos mais populares para o ajuste de modelos de regressão não linear, o chamado Método de Gauss-Newton. Embora as linguagens computacionais não façam uso de algoritmos exatamente coincidentes com os passos aqui descritos, a apresentação de seus fundamentos poderá fornecer uma apreciação acerca das dificuldades numéricas para a obtenção das estimativas dos parâmetros dos modelos não-lineares.

2.1 Método de Gauss-Newton

Um dos métodos mais utilizados para o ajuste de modelos de regressão não linear é o Método de Gauss-Newton. Conforme mencionado anteriormente, a sequência de passos a seguir não necessariamente corresponde exatamente àquela utilizada por ferramentas computacionais, como a linguagem **R**, que em geral implementam rotinas adicionais de otimização numérica. Maiores detalhes sobre este método podem ser encontrados, por exemplo, nas publicações de AGUIAR (2012) e de SILVA *et al.* (2019).

O Método de Gauss-Newton consiste em se fazer uma aproximação em série de Taylor da relação $f(\boldsymbol{\xi}_i, \boldsymbol{\beta})$, utilizando um polinômio de primeira ordem, em uma vizinhança $\boldsymbol{\beta} = \boldsymbol{\beta}^0$. Esta vizinhança corresponde a um valor inicial, uma primeira estimativa, dos elementos de $\boldsymbol{\beta}$. Desta maneira, tem-se agora um modelo linearizado, e assim pode-se obter uma nova estimativa de $\boldsymbol{\beta}$, utilizando o método dos quadrados mínimos, em uma expressão análoga à 2.4. Esta nova estimativa passa a corresponder à nova vizinhança, e o processo pode ser repetido. Assim procedendo de forma iterativa, caso haja convergência, obtém-se a estimativa final de $\boldsymbol{\beta}$.

Utilizando a aproximação por série de Taylor no modelo 2.5, até o primeiro grau, tem-se que, em uma vizinhança $\boldsymbol{\beta} = \boldsymbol{\beta}^0$:

$$\begin{aligned} Y_i \approx f(\boldsymbol{\xi}_i, \boldsymbol{\beta}^0) + \left.\frac{\partial f(\boldsymbol{\xi}_i, \boldsymbol{\beta})}{\partial \beta_1}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} (\beta_1 - \beta_1^0) + \left.\frac{\partial f(\boldsymbol{\xi}_i, \boldsymbol{\beta})}{\partial \beta_2}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} (\beta_2 - \beta_2^0) \\ + \ldots + \left.\frac{\partial f(\boldsymbol{\xi}_i, \boldsymbol{\beta})}{\partial \beta_p}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^0} (\beta_p - \beta_p^0) + \epsilon_i \end{aligned} \tag{2.6}$$

O modelo 2.6 também pode ser expresso matricialmente, através de:

$$\mathbf{y} = \mathbf{F}(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\epsilon}$$

em que o vetor $\mathbf{y}$ possui elementos:

$$Y_i - f(\boldsymbol{\xi}_i, \boldsymbol{\beta}^0) \tag{2.7}$$

e $\mathbf{F}$ é a matriz de derivadas primeiras de $f(\boldsymbol{\xi}_i, \boldsymbol{\beta})$ em relação aos elementos de $\boldsymbol{\beta}$, avaliadas em $\boldsymbol{\beta} = \boldsymbol{\beta}^0$. Desta maneira, o modelo está linearizado. Utilizando uma expressão semelhante a 2.4, tem-se:

$$\mathbf{F}'\mathbf{F}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = \mathbf{F}'\mathbf{y}$$

A solução $\hat{\boldsymbol{\beta}}$ deste sistema pode ser interpretada como sendo o valor de uma primeira

iteração, podendo ser representada por $\boldsymbol{\beta}^1$. Assim:

$$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^0 + (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y}$$

E assim, de forma recursiva, tem-se que, na iteração $j+1$:

$$\boldsymbol{\beta}^{j+1} = \boldsymbol{\beta}^j + (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{y} \tag{2.8}$$

Assim procede-se iterativamente, até que se atinja convergência. Este procedimento recursivo corresponde ao método de Gauss-Newton.

2.1.1 Um Exemplo

O algoritmo de Gauss-Newton será a seguir ilustrado, considerando um pequeno conjunto de dados, que foram ajustados ao seguinte modelo não-linear:

$$Y_i = \beta_1 + X_i^{\beta 2} + \epsilon_i \tag{2.9}$$

o qual é um caso particular de 2.5, em que $f(\boldsymbol{\xi}_i, \boldsymbol{\beta}) = \beta_1 + X_i^{\beta 2}$. Trata-se de um modelo não-linear, pois:

$$\frac{\partial f(\boldsymbol{\xi}_i, \boldsymbol{\beta})}{\partial \beta_2} = X_i^{\beta 2} \ln X_i$$

.

Um pequeno conjunto de dados simulado para ilustrar o método está apresentado na Tabela 1. Tratam-se de dados simulados, que poderiam representar, por exemplo, o tempo de exposição ao sol de uma planta e seu crescimento, em *cm*.

Tabela 1: Conjunto de dados simulado, referente a uma variável resposta Y e uma variável expllicativa X.

X	Y
2	4,9
4	9,3
6	16,7
8	24,3

Antes de mais nada, é preciso definir a matriz $\mathbf{F}$ de derivadas primeiras, de $f(\boldsymbol{\xi}_i, \boldsymbol{\beta})$ em relação aos parâmetros:

$$\mathbf{F} = \begin{bmatrix} 1 & 2^{\beta_2} \ln 2 \\ 1 & 4^{\beta_2} \ln 4 \\ 1 & 6^{\beta_2} \ln 6 \\ 1 & 8^{\beta_2} \ln 8 \end{bmatrix} \tag{2.10}$$

O vetor de parâmetros é dado por:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{2.11}$$

e o vetor $\mathbf{y}$, conforme definido em 2.7, é dado por:

$$\mathbf{y} = \begin{bmatrix} 4,9 - \beta_1 - 2^{\beta_2} \\ 9,3 - \beta_1 - 4^{\beta_2} \\ 16,7 - \beta_1 - 6^{\beta_2} \\ 24,3 - \beta_1 - 8^{\beta_2} \end{bmatrix} \tag{2.12}$$

Partindo, por exemplo, dos chutes iniciais $\beta_1^0 = 1,8$ e $\beta_2^0 = 2$, isto promove uma primeira composição da matriz $\mathbf{F}$ e do vetor $\mathbf{y}$, que por sua vez produzem as estimativas da primeira iteração. Prosseguindo assim iterativamente, tem-se sucessivas atualizações das estimativas. Os seus valores, para as 6 primeiras iterações, estão apresentadas na Tabela 2.

Tabela 2: Estimativas dos parâmetros de um modelo de regressão não linear, ao longo de 6 iterações, utilizando o método de Gauss-Newton.

Iteração	$\hat{\beta}_1$	$\hat{\beta}_2$
0	1,8	2,0
1	0,892154	1,692782
2	1,677591	1,533818
3	1,800792	1,499355
4	1,805532	1,497974
5	1,805566	1,497971
6	1,805566	1,497971

Percebe-se, pela Tabela 2, que com apenas 6 iterações obteve-se convergência, até a sexta casa decimal.

2.2 Regressão Logística

O termo "Regressão Logística" talvez seja mais conhecido no contexto da teoria de Modelos Lineares Generalizados (ver, por exemplo, AGRESTI, 2002), em que se tem a variável dependente como categórica, ou seja, em que as respostas correspondem (por exemplo) a diferentes gêneros, temas, tipos *etc* ou mesmo realizações de uma variável resposta binária, tais como sim ou não, feminino ou masculino. A representação gráfica desse modelo é uma curva sigmoide, semelhante ao formato da letra S, com uma imagem da função correspondendo ao intervalo entre 0 e 1. Ou seja, trata-se de um enfoque em que probabilidades de pertencimento a uma determinada categoria são expressas como funções de variáveis explicativas X.

Contudo, no presente trabalho, serão considerados modelos de regressão logística com esta tendência sigmoide, mas admitindo que a variável resposta é contínua e com distribuição normal, ou mais precisamente, que os resíduos deste modelo tenham distribuição normal com média 0 e variância σ^2 (DRAPER & SMITH, 2014). Trata-se de um modelo não-linear, como se pode observar pelo seu modelo estatístico:

$$Y_i = \frac{\beta_3}{1 + e^{-(\beta_1 + \beta_2 X_i)}} + \epsilon_i \tag{2.13}$$

em que os parâmetros β_2 e β_3 são positivos.

O gráfico desta função pode ser apreciado na Figura 3.1.

A regressão logística é utilizada em diversas áreas do conhecimento, como em Econometria, mas, dada a sua natureza sigmoide, é especialmente empregada para descrever fenômenos de crescimento, seja populacional (populações, colônias *etc*), seja individual (peso ou altura de organismos). Nestes casos, é muito comum que a variável X corresponda ao tempo, embora possa eventualmente corresponder a alguma outra variável ambiental.

2.2.1 Propriedades

Nesta subseção serão apresentadas algumas propriedades da regressão logística, de maneira a propiciar maior compreensão do modelo, bem como auxiliar na interpretação de seus parâmetros.

Um dos aspectos mais evidentes do gráfico da função é a existência de duas assíntotas

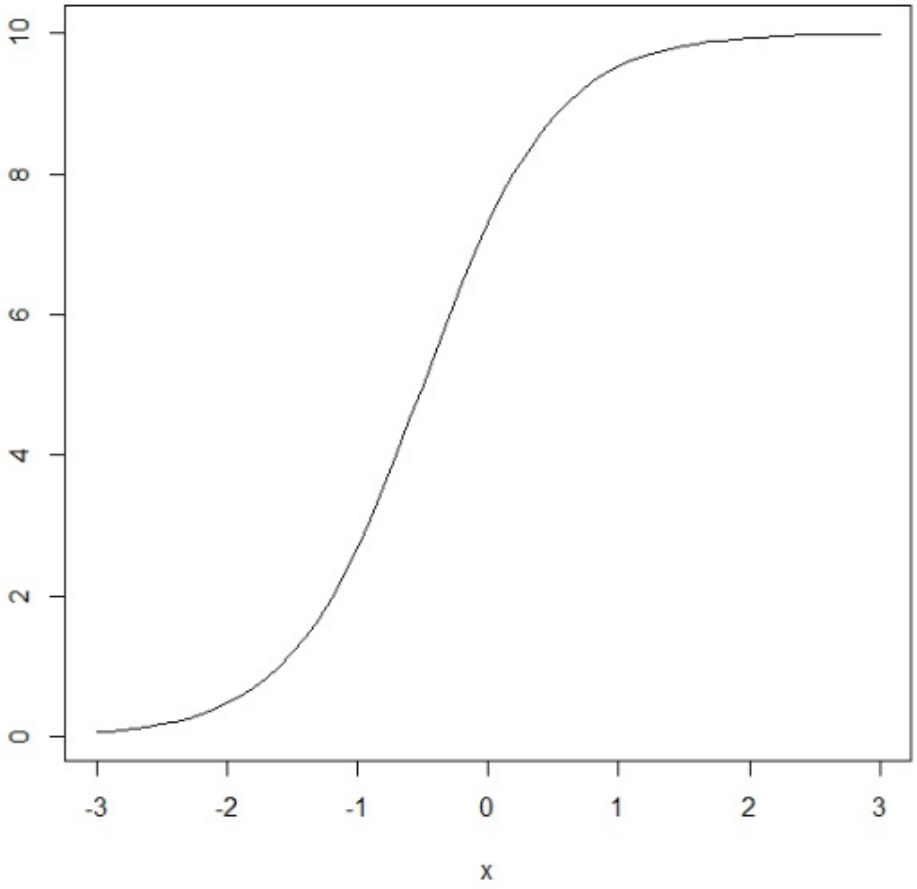

Figura 1: Exemplo de um modelo de regressão logística.

bem definidas. Considere-se a parte determinística do modelo, ou seja:

$$f(x, \boldsymbol{\beta}) = \frac{\beta_3}{1 + e^{-(\beta_1 + \beta_2 x)}}$$

Sendo o parâmetro β_2 positivo (grande maioria das situações), então pode-se verificar que:

$$\lim_{x \to \infty} f(x, \boldsymbol{\beta}) = \beta_3 \tag{2.14a}$$

$$\lim_{x \to -\infty} f(x, \boldsymbol{\beta}) = 0 \tag{2.14b}$$

Ou seja, o parâmetro β_3 pode ser considerado o valor máximo para o qual a variável resposta tende, à medida que X aumenta. Quando esta tende ao menos infinito, a variável resposta tende a zero.

Uma das possíveis generalizações do modelo de regressão logística é a de considerar que o parâmetro β_2 seja negativo. A curva continuará sigmoide, mas com um formato de

S "invertido". Neste caso, os resultados acima se invertem. A variável resposta tenderá a zero, à medida que X aumenta, e tenderá a β_3 à medida que X diminui.

O ponto de inflexão da curva ocorre para o seguinte valor de X:

$$x = \frac{-\beta_1}{\beta_2} \tag{2.15}$$

O ponto de inflexão corresponde ao ponto em que o modelo tem sua inclinação máxima. Se o modelo estiver descrevendo um fenômeno de crescimento, este ponto corresponderá ao tempo em que o crescimento acontece com velocidade máxima. Ou seja, trata-se do ponto de máximo da derivada primeira. Assim, o ponto de inflexão acima foi obtido igualando a derivada segunda de $f(x, \boldsymbol{\beta})$ em relação a x a zero. Após obtidas as estimativas dos parâmetros, em geral há interesse prático em se obter este ponto de inflexão da curva.

Por se tratar de um modelo de regressão não-linear, faz-se necessário o uso de um método numérico para o seu ajustamento. Caso se utilize o método de Gauss-Newton, são necessárias as derivadas primeiras de $f(x, \boldsymbol{\beta})$ em relação aos parâmetros do modelo, as quais estão apresentadas abaixo:

$$\frac{\partial}{\partial \beta_1} f(x, \boldsymbol{\beta}) = \frac{\beta_3 e^{-(\beta_1 + \beta_2 x)}}{\left[1 + e^{-(\beta_1 + \beta_2 x)}\right]^2} \tag{2.16a}$$

$$\frac{\partial}{\partial \beta_2} f(x, \boldsymbol{\beta}) = \frac{\beta_3 x e^{-(\beta_1 + \beta_2 x)}}{\left[1 + e^{-(\beta_1 + \beta_2 x)}\right]^2} \tag{2.16b}$$

$$\frac{\partial}{\partial \beta_3} f(x, \boldsymbol{\beta}) = \frac{1}{1 + e^{-(\beta_1 + \beta_2 x)}} \tag{2.16c}$$

2.3 Modelos Segmentados

Outra classe de modelos não-lineares de grande potencial de aplicações e versatilidade consiste a dos chamados *modelos segmentados*. Essencialmente, em tais modelos o domínio da parte estrutural do modelo é subdividido em segmentos, de maneira a permitir diferentes relações funcionais entre X e Y (HOFFMANN, 2016).

Para uma melhor compreensão, esta classe de modelos será inicialmente descrita em sua versão linear, para posterior generalização para a versão não-linear.

Conforme apontado acima, nos modelos segmentados há diferentes relações entre X

e Y, conforme o segmento. Didaticamente, estes modelos poderiam ser classificados em duas categorias, quais sejam, modelos descontínuos ou contínuos. Na primeira não há uma continuidade entre os dois segmentos, ao passo que na segunda há um mesmo valor de Y no valor de X que separa os dois segmentos, ou seja, há uma continuidade entre os segmentos.

2.3.1 Modelo Descontínuo

Para uma melhor visualização acerca dos modelos segmentados descontínuos, suponha-se que a variável dependente seja o preço de algum produto, e a independente o tempo (em meses). Suponha-se ainda que nos quatro primeiros meses o preço aumentou gradativamente, mas que, entre o quarto e o quinto meses, por algum motivo (por exemplo alguma intervenção governamental) o preço do produto sofreu uma queda, a partir da qual o preço continuou a variar em função do tempo, mas com outra tendência de inclinação. O exemplo de um conjunto de dados possível está apresentado na Tabela 3.

Tabela 3: Dados (fictícios) do preço de um produto em função do tempo, em meses.

Mês	Preço
1	1,5
2	3,1
3	7,5
4	10,4
5	2,4
6	3,5
7	5,0
8	5,2

Em uma situação como essa, a intervenção ocorrida entre os meses 4 e 5 caracteriza uma descontinuidade, separando o domínio da função em dois segmentos: até o mês 4, e outro a partir do mês 5. Suponha-se que se esteja admitindo que em ambos os segmentos a relação funcional entre X e Y seja uma reta, cada qual com diferentes intercepto e coeficiente angular. Este comportamento pode ser visualizado na Figura 2.

Em princípio, poder-se-ia pensar em ajustar estas duas retas em separado, tomando-se cada segmento isoladamente. Entretanto, esta abordagem teria a desvantagem da redução do tamanho da amostra, uma vez que a cada segmento corresponderia um conjunto de dados em separado, o que reduziria o número de graus de liberdade para o resíduo e a precisão das estimativas.

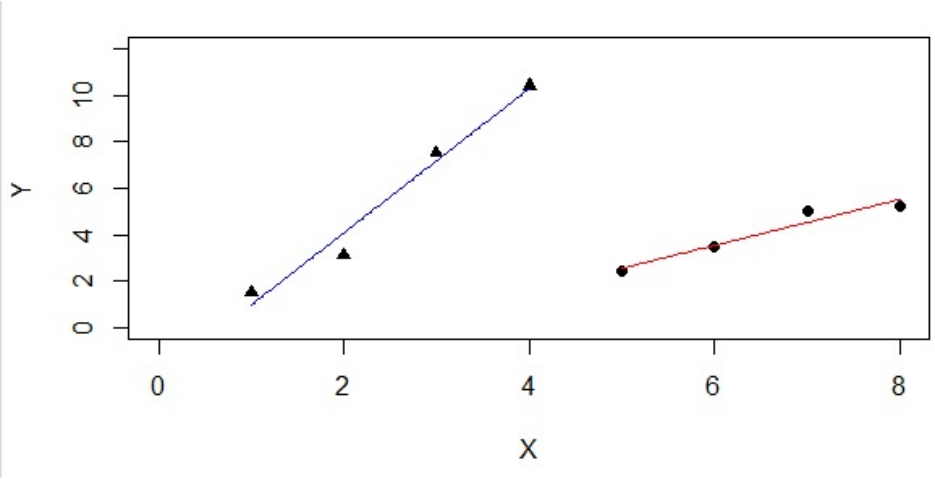

Figura 2: Exemplo de um modelo segmentado descontínuo.

Este problema poderia ser contornado com o ajustamento simultâneo dos dois segmentos de reta, através da utilização de variáveis indicadoras, também conhecidas como variáveis *dummy*, conforme a terminologia inglesa (DRAPER & SMITH, 2014; HOFFMANN, 2016). O procedimento consiste na criação de uma variável binária, que assume um valor para o segmento 1, e um outro valor para o segmento 2. Desta forma, pode-se fazer o ajuste das duas retas a partir de um único modelo estatístico.

Sem perda de generalidade, esta variável *dummy* será representada pela letra Z, admitindo-se que assuma o valor 0 para o primeiro segmento, e o valor 1 para o segundo segmento. Desta maneira, o modelo estatístico poderia ser dado por:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \epsilon_i \tag{2.17}$$

Em posse de tal modelo, tem-se que, no primeiro segmento, dado que $Z = 0$, a relação funcional entre X e Y será dada por:

$$Y_i = \beta_0 + \beta_2 X_i$$

correspondente a uma reta com intercepto β_0, e coeficiente angular β_2. Já no segundo segmento, uma vez que $Z = 1$, tem-se:

$$Y_i = \beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i$$

ou seja:

$$Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i$$

a qual corresponde a uma reta com intercepto $\beta_0 + \beta_1$, e coeficiente angular $\beta_2 + \beta_3$.

Caso houvesse mais do que 2 segmentos, novas variáveis *dummy* seriam necessárias. Para uma situação com k segmentos, seriam necessárias $k-1$ variáveis *dummy* para discriminá-los.

Aqui cabe uma observação. Ao se utilizar a abordagem do uso de variáveis *dummy*, em relação aos ajustes em separado para cada segmento, tem-se que as estimativas dos parâmetros são as mesmas. Contudo, por se tratar de um ajuste simultâneo, tem-se um número maior de graus de liberdade para o resíduo, e assim uma maior precisão (menores erros padrões) e maior poder para testes de hipóteses. As estimativas do intercepto e do coeficiente angular estão apresentadas na Tabela 4.

Tabela 4: Estimativas do intercepto e da inclinação das retas de dois segmentos, em um modelo segmentado descontínuo (dados fictícios).

Segmento	Intercepto	Inclinação
1	-2,15	3,11
2	-2,41	0,99

Uma outra parametrização, que conduz aos mesmos resultados, consiste em explicitar o intercepto e o coeficiente angular de cada segmento. Esta parametrização alternativa é apresentada aqui, pois irá simplificar a obtenção do modelo estatístico em outras categorias de modelos segmentados. Assim, denominando-se os dois segmentos do domínio de X por A e B poder-se-ia definir:

$$Y = \begin{cases} \beta_0 + \beta_1 X, \text{ se } X \in A \\ \gamma_0 + \gamma_1 X, \text{ se } X \in B \end{cases} \tag{2.18}$$

Aqui, apenas por simplicidade de apresentação, está-se omitindo o termo residual. A variável *dummy* é definida da mesma maneira, ou seja:

$$Z = \begin{cases} 0, \text{ se } X \in A \\ 1, \text{ se } X \in B \end{cases}$$

Para escrevermos o modelo 2.18 em uma única expressão, será feito aqui um artifício. Os termos da reta do segmento A serão somados e subtraídos ao segmento B. Assim, o modelo 2.18 pode alternativamente ser expresso como:

$$Y = \begin{cases} \beta_0 + \beta_1 X, \text{ se } X \in A \\ \gamma_0 + \gamma_1 X + \beta_0 - \beta_0 + \beta_1 X - \beta_1 X, \text{ se } X \in B \end{cases}$$

Desta maneira, a função acima pode ser expressa utilizando uma única relação, com o auxílio da variável *dummy*:

$$Y = \beta_0 + \beta_1 X + (\gamma_0 + \gamma_1 X - \beta_0 - \beta_1 X)Z \tag{2.19}$$

Pode-se observar em 2.19 que, para o primeiro segmento ($Z = 0$) tem-se:

$$Y = \beta_0 + \beta_1 X$$

enquanto que, para o segundo segmento ($Z = 1$), tem-se:

$$Y = \gamma_0 + \gamma_1 X$$

conforme definido em 2.18.

2.3.2 Modelo Contínuo, Interseção Conhecida

Aqui será considerada a situação em que os dois segmentos apresentam uma continuidade, ou seja, para o valor de X que separa os dois segmentos, tem-se um mesmo valor de Y. Quando este valor de X é conhecido, diz-se que a interseção é conhecida, e assim o modelo segmentado ainda é de natureza linear.

Como exemplo de motivação, considere-se que o colesterol de uma pessoa está sendo acompanhado e medido ao longo de diferentes meses, com tendência de aumento. Considere-se ainda que entre os meses 4 e 5 o indivíduo começou a praticar esportes, o que começa a provocar uma diminuição no nível de colesterol, nos meses subsequentes. Para um indivíduo, o nível de colesterol aumenta ou diminui de maneira contínua, não fazendo sentido uma descontinuidade, como na situação anterior. Este exemplo pode ser visualizado na Figura 3.

Suponha-se que se conheça o valor de X (que será representado aqui por θ) que separa os dois segmentos. Por exemplo, se o indivíduo começou a praticar esportes exatamente aos 15 dias entre os meses 4 e 5, então se teria $\theta = 4,5$.

Em uma situação como essa, e ainda admitindo-se que também se trata de uma reta

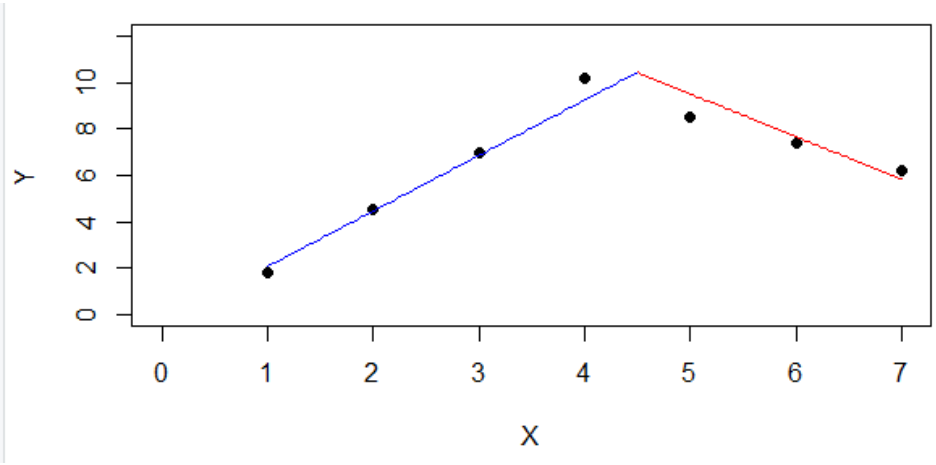

Figura 3: Exemplo de um modelo segmentado contínuo com θ conhecido.

em cada segmento, poder-se-ia escrever:

$$Y = \begin{cases} \beta_0 + \beta_1 X, \text{ se } X < \theta \\ \gamma_0 + \gamma_1 X, \text{ se } X > \theta \\ \beta_0 + \beta_1 X = \gamma_0 + \gamma_1 X, \text{ se } X = \theta \end{cases} \tag{2.20}$$

A terceira das relações em 2.20 é que garante a continuidade da função. Ou seja, quando $X = \theta$, tem-se que:

$$\beta_0 + \beta_1 \theta = \gamma_0 + \gamma_1 \theta$$

e assim pode-se colocar um dos parâmetros em evidência, escrevendo-o como função dos demais. Por exemplo, pode-se escrever:

$$\gamma_0 = \beta_0 + \beta_1 \theta - \gamma_1 \theta \tag{2.21}$$

Esta dependência entre os parâmetros decorre da continuidade. Da mesma maneira como foi feito no exemplo anterior, o modelo também pode ser escrito da mesma maneira como feito em 2.19:

$$Y = \beta_0 + \beta_1 X + (\gamma_0 + \gamma_1 X - \beta_0 - \beta_1 X) Z$$

com a diferença de que agora há uma dependência entre os parâmetros. Substituindo 2.21 em 2.19, resulta:

$$Y = \beta_0 + \beta_1 (X + Z\theta - XZ) + \gamma_1 (XZ - \theta Z) \tag{2.22}$$

Sendo θ conhecido, os termos dentro dos parênteses de 2.22 também são, e assim verifica-

se que o modelo é linear. Este modelo possui 3 parâmetros a serem estimados (e não 4, pois, por exemplo, γ_0 é função dos demais), ao contrário do exemplo anterior, em que havia 4 parâmetros a serem estimados.

2.3.3 Modelo Contínuo, Interseção Desconhecida

Finalmente, na última categoria de modelos segmentados, tem-se a situação em que θ é desconhecido. Considerando ainda a situação com dois segmentos, sendo uma reta em cada segmento, o modelo é o mesmo que 2.22, com a diferença de que θ agora é um parâmetro a ser estimado. Com isso, o modelo passa a ser não-linear. Isto é facilmente percebido, bastando observar que, em 2.22, há termos que envolvem produtos de parâmetros, como $\beta_1\theta$ e $\gamma_1\theta$, quando $Z = 1$.

Nesta abordagem, embora θ seja desconhecido, é importante notar que se admite que se saiba entre quais valores de X o parâmetro θ está localizado. Isto porque se está admitindo que a variável *dummy* Z é conhecida, ou seja, sabe-se exatamente quais pontos pertencem a cada um dos segmentos.

3 Material e Métodos

Neste trabalho, o uso de regressão não linear é ilustrado utilizando dois bancos de dados referentes a características de leite materno, gentilmente disponibilizados pela Profa. Camila Carvalho Menezes, do Departamento de Nutrição Clínica e Social, da Escola de Nutrição da Universidade Federal de Ouro Preto. Tais dados foram utilizados para o ajuste de uma regressão logística e de um modelo segmentado, ambos com modificações sugeridas para melhoria de ajuste, em relação aos modelos conforme propostos originalmente. No caso da regressão logística, é proposta a inclusão de um parâmetro adicional (uma constante), de maneira a flexibilizar as assíntotas da curva, e no caso do modelo segmentado foram considerados dois segmentos, sendo que em um deles a relação entre a variável preditora e a variável resposta correspondeu a uma curva, e não uma reta.

Os dados utilizados para o ajuste da regressão logística estão descritos em PARREIRAS *et al.* (2020), enquanto que a descrição do conjunto de dados utilizado no ajuste do modelo segmentado pode ser encontrada em NOGUEIRA (2020). Ambos os estudos trataram de avaliar características do leite materno, cada qual considerando diferentes aspectos. Estes estudos são brevemente descritos a seguir, para uma melhor compreensão acerca da natureza dos dados considerados.

O leite materno traz consigo tudo que um bebê precisa para sua alimentação primária e ajuda todo o seu desenvolvimento. Eventualmente, a mãe da criança não reúne condições de amamentá-lo, e nestes casos uma alternativa corresponde a amamentar os recém-nascidos com leite materno doado por mulheres em puerpério.

O leite doado sempre é processado para evitar infecções como as das bactérias *Escherichia coli*, *Staphylococcus aureus* e *Salmonella ssp.* O processamento padrão consiste em uma pasteurização lenta, que corresponde a super aquecer o produto lentamente, sendo então posteriormente totalmente refrigerado e estocado. Este tratamento térmico convencional por vezes tem sido questionado, dada a possibilidade de remoção de componentes antioxidantes essenciais e lipídios para o bebê. Desta forma, os estudos de PARREIRAS

et al. (2020) e NOGUEIRA (2020) se inserem em uma linha de pesquisa que investiga outros métodos de processamento, como o da ultrassonificação.

3.1 Coleta do Leite Humano

Os conjuntos de dados de ambos os estudos têm a mesma origem, sendo dados de doadoras recorrentes, com idade superior a 20 anos, do Banco de Leite Humano (BLH) da Santa Casa da Misericórdia de Ouro Preto. A coleta foi feita de forma manual, em casa, sendo depois o leite refrigerado. Esta coleta ocorreu entre novembro de 2018 e janeiro de 2019, e serviu de base para duas pesquisas, considerando diferentes métodos de processamento em cada uma, e eventualmente outros fatores, como tempo de armazenamento e temperatura. No trabalho de PARREIRAS *et al.* (2020), houve o interesse em se estudar a presença de microrganismos, enquanto que o de NOGUEIRA (2020) investigou características lipídicas do leite.

3.2 Micro-Organismos no Leite Humano

Com o propósito de ilustrar a regressão logística, foram utilizados os dados de PARREIRAS *et al.* (2020), que estudou dois métodos de processamento: método térmico, e termossonificação. A variável resposta correspondeu ao logaritmo (na base 10) do número de bactérias das espécies: *Escherichia coli*, *Staphylococcus aureus* e *Salmonella ssp*, a qual foi relacionada com uma variável preditora, correspondente à temperatura utilizada.

Originalmente, este experimento foi instalado considerando o modelo de uma superfície de resposta, uma vez que havia o interesse em se modelar a variável resposta tanto em função da temperatura como também do tempo. Como este último acabou por se mostrar um fator não relevante, não foi considerado aqui. De qualquer maneira, o delineamento utilizado foi o composto central rotacional (DCCR). Neste delineamento, assim como em tantos outros propostos na teoria de superfícies de resposta, o número de repetições por tratamento varia, de maneira a maximizar a precisão das estimativas, ao mesmo tempo economizando recursos.

No presente trabalho, para cada método de processamento, foi ajustada uma curva de regressão logística, com a variável resposta em função da temperatura. Aqui foi proposta uma parametrização alternativa, pela inclusão, ao modelo 2.13, de um parâmetro adicional β_0, correspondente a uma constante:

$$Y_i = \beta_0 + \frac{\beta_3}{1 + e^{-(\beta_1 + \beta_2 X_i)}} + \epsilon_i \tag{3.1}$$

A vantagem desta parametrização consiste em uma flexibilização das assíntotas da curva. Enquanto que no modelo 2.13 tais assíntotas ocorrem em $Y = 0$ e $Y = \beta_3$, no modelo 3.1 estas correspondem a β_0 e $\beta_0 + \beta_3$. Outra flexibilização importante consistiu em se admitir a possibilidade de o parâmetro β_2 ser negativo, viabilizando o ajuste de curvas no formato de "S invertido".

3.3 Lipídios no Leite Humano

Para a ilustração de um modelo bisegmentado, foram utilizados os dados do trabalho de NOGUEIRA (2020). Nesse estudo, as amostras de leite foram congeladas a temperatura inferior a -18^{o}C, sendo avaliadas em relação ao teor de diferentes tipos de lipídios, ao longo de 0, 15, 30, 60 e 120 dias de armazenamento. É importante ressaltar que diferentes amostras eram avaliadas em cada tempo. Ou seja, não se voltava a uma mesma amostra nos diferentes tempos, o que poderia acarretar dependência residual entre observações, ferindo o princípio da independência. Cada observação correspondeu a uma alíquota de 5 mL de leite humano aquecida em banho maria a 85^{o}C, por 45 minutos.

O estudo utilizou o delineamento inteiramente casualizado (DIC) com três repetições. Os tratamentos estavam arranjados em uma estrutura fatorial, correspondendo a combinações entre métodos de processamento e tempos de armazenamento.

Neste trabalho, o lipídio considerado foi o hexanal, identificado como o principal aldeído volátil vindo do leite materno, sendo um indicador químico sensível e útil para a avaliação das reações de oxidação no leite (NOGUEIRA, 2020). Para um dos métodos de processamento, esta variável resposta apresentou um comportamento que possibilitou o ajuste de um modelo bisegmentado interessante, no qual a relação funcional entre Y e X correspondeu a uma reta em um segmento, e a uma curva, no outro segmento.

4 Resultados e Discussão

4.1 Regressão Logística

Foi realizado o ajuste de seis modelos de regressão logística, tendo o logaritmo do número de microrganismos por alíquota de leite, em função da temperatura. Foram três os microrganismos considerados (*Escherichia coli*, *Staphylococcus aureus* e *Salmonella ssp*), e dois métodos de processamento do leite materno: tratamento térmico (pasteurização) e termossonicação (ultrassom), totalizando assim as seis situações consideradas.

Em um primeiro momento, será discutido com maior detalhe o ajuste do modelo aos dados de *Escherichia coli*, para o método térmico, e em seguida serão apresentados os demais ajustes.

4.1.1 *Escherichia coli*

A Tabela 5 apresenta os dados referentes à quantidade de microrganismos *Escherichia coli* por alíquota de leite (Y), em função da temperatura de processamento. A variável resposta correspondeu ao logaritmo (na base 10) da quantidade de microrganismos por alíquota.

Conforme se pode observar na Tabela 5, há valores repetidos da variável preditora (temperatura), com números de repetições estabelecidos conforme o delineamento composto central rotacional (DCCR), utilizado neste estudo. No ajuste de modelos de regressão linear, o fato de haver repetições propicia a decomposição da variação residual em duas fontes: o chamado "Erro Puro", e a "Falta de Ajustamento". O Erro Puro corresponde à variação entre repetições de um mesmo valor de X, enquanto que a Falta de Ajustamento (em inglês, *lack of fit*) quantifica a variação entre valores de X que não é explicada pelo modelo de regressão. Em uma regressão linear, testar a significância da Falta de Ajustamento é um procedimento interessante, pois possibilita verificar se o modelo de regressão escolhido é apropriado. Mais adiante, será visto que este enfoque também pôde

Tabela 5: Dados referentes ao logaritmo do número de microrganismos *Escherichia coli* por alíquota de leite (Y), em função da temperatura de processamento.

Temperatura	Y
31	7,04
36	6,99
36	6,97
48	6,96
48	6,94
48	6,96
48	6,88
48	6,96
60	4,34
60	4,20
65	4,07

ser adaptado para o modelo de regressão não linear considerado aqui.

Para o ajuste da regressão logística, conforme mencionado no Capítulo anterior, foram consideradas duas modificações: a inclusão de um termo constante no modelo (conforme se pode observar no modelo 3.1), bem como a possibilidade de o parâmetro β_2 ser negativo.

Um aspecto importante do ajustamento é a questão da determinação dos chutes iniciais para os parâmetros, embora eventualmente haja programas computacionais de análise com rotinas integradas para tal. Na presente situação, foi utilizado um procedimento para obtenção de chutes iniciais razoáveis, conforme descrito a seguir.

A obtenção de chutes iniciais basicamente consistiu em uma linearização da parte estrutural do modelo de regressão logística. Por se tratar de um critério de obtenção de valores aproximados, em um primeiro momento foi ignorado o termo constante do modelo (β_0). Assim, partiu-se da seguinte relação estrutural (também sem o termo residual):

$$Y = \frac{\beta_3}{1 + e^{-(\beta_1 + \beta_2 X)}}$$

a qual foi linearizada conforme os passos a seguir:

$$\frac{Y}{\beta_3} = \frac{1}{1 + e^{-(\beta_1 + \beta_2 X)}}$$

$$\frac{\beta_3}{Y} = 1 + e^{-(\beta_1 + \beta_2 X)}$$

$$\frac{\beta_3}{Y} - 1 = e^{-(\beta_1 + \beta_2 X)}$$

$$\frac{\beta_3 - Y}{Y} = e^{-(\beta_1 + \beta_2 X)}$$

$$\frac{Y}{\beta_3 - Y} = e^{\beta_1 + \beta_2 X}$$

$$\frac{Y/\beta_3}{1 - Y/\beta_3} = e^{\beta_1 + \beta_2 X}$$

e assim:

$$\ln\left(\frac{Y/\beta_3}{1 - Y/\beta_3}\right) = \beta_1 + \beta_2 X \tag{4.1}$$

sendo que esta última transformação corresponde à chamada *função logito*. Dessa maneira, tendo-se um chute inicial para β_3, tem-se que podemos ajustar 4.1 como um modelo de regressão linear simples.

Nesta parametrização ignorando o termo constante, o parâmetro β_3 corresponde ao valor da assíntota superior. Assim, pode-se tomar como um chute inicial seu algum um valor próximo ao valor máximo observado na amostra. Com isso, ao ajustar a regressão linear simples, obtém-se os chutes iniciais para β_1 e β_2. Resta obter um chute inicial para β_0, que corresponde à assíntota inferior. Pode-se assim tomar um valor próximo do menor valor observado na amostra. Procedendo desta maneira, obteve-se o seguinte modelo ajustado:

$$\hat{Y}_i = 4,05 + \frac{2,95}{1 + e^{-(29,50 - 0,53X_i)}}$$

sendo que o gráfico correspondente pode ser observado na Figura 4.

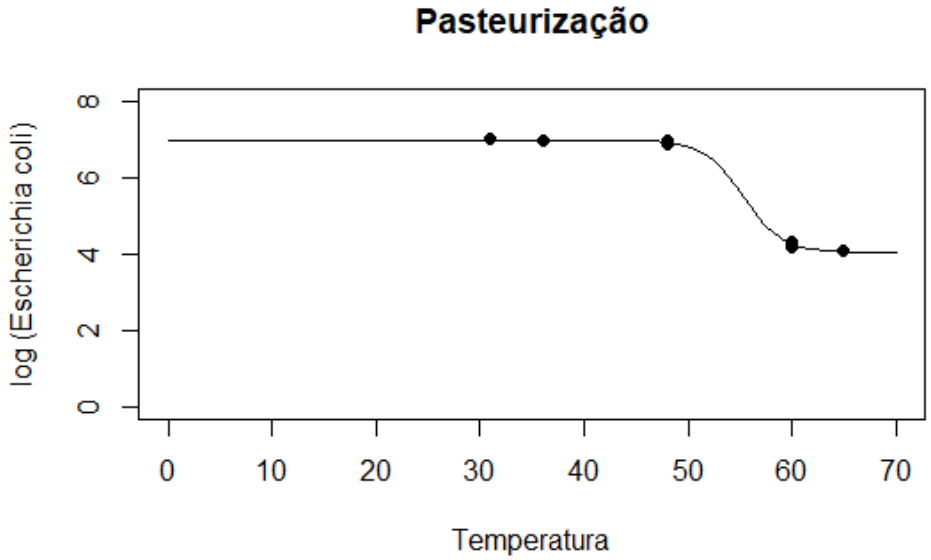

Figura 4: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Escherichia coli* em alíquotas de leite, em função da temperatura de processamento (método: pasteurização).

Verifica-se, na Figura 4, como foi interessante a flexibilização quanto ao sinal de β_2. Como aqui foi negativo, tem-se uma curva em formato de "S invertido". Ou seja, à me-

dida que aumenta a temperatura de processamento, a quantidade de microrganismos vai tendendo a diminuir.

Para verificar se o modelo foi significativo e apropriado, é conveniente realizar uma Análise de Variância (ANOVA). Esta técnica, em princípio, se aplica a modelos lineares, contendo termo constante, trabalhando-se assim com somas de quadrados ajustadas (corrigidas) para a constante do modelo. Como o presente modelo de regressão logística apresenta um termo constante (β_0), é possível testar a significância do modelo ajustado para β_0, bem como os desvios de regressão. Esta ANOVA está apresentada na Tabela 6. Perceba que o número de graus de liberdade para regressão é igual a 3 (e não 4), pois está ajustada para β_0. Como há 5 valores distintos de temperatura (4 graus de liberdade), então sobra $4 - 3 = 1$ grau de liberdade para a Falta de Ajustamento (também chamada de Desvios de Regressão). Com base nas estatísticas F desta ANOVA, verifica-se que a regressão foi altamente significativa, enquanto que a Falta de Ajustamento não foi (Tabela 6). Isto indica que o modelo escolhido foi adequado, ou seja, a variação remanescente, não explicada pelo modelo de regressão, não foi significativa.

Tabela 6: Análise de variância do ajuste de um modelo de regressão logística a dados referentes à presença de *Escherichia coli* em alíquotas de leite, em função da temperatura de processamento.

Causas de Variação	GL	SQ	QM	F
Regressão	3	16,6436	5,5479	2249,14***
Desvios (Falta de Ajustamento)	1	0,0024	0,0024	0,97
Erro Puro	6	0,0148	0,0025	—
Total	10	16,6608	—	—

*** significativo a 0,1% de probabilidade.

Caso haja interesse na realização de testes de hipóteses e/ou intervalos de confiança para cada parâmetro individualmente, pode-se utilizar a aproximação normal, considerando-se os erros padrões de cada estimador. Estes erros padrões podem ser obtidos multiplicando-se a inversa da matriz $\mathbf{F}'\mathbf{F}$ pela estimativa de σ^2 (aqui, correspondente ao quadrado médio do Erro Puro). Na diagonal desta matriz, tem-se as variâncias de cada estimador. Extraindo-se a raiz quadrada destes elementos, tem-se os erros padrões, os quais podem ser utilizados, por exemplo, para construir intervalos de confiança. Este procedimento, contudo, não será apresentado aqui.

É interessante notar que o ajuste deste modelo de regressão logística permite realizar uma série de inferências de interesse prático. Por exemplo, a assíntota superior é estimada como tendo valor igual a $4,05+2,95=7,00$, e a assíntota inferior é estimada como sendo 4,05. Ou seja, é interessante notar que o método de processamento térmico do leite tende a não eliminar totalmente os microrganismos do leite (*Escherichia coli*), permanecendo uma quantidade residual, estimada como sendo da ordem de $10^{4,05}=11.220$ microrganismos por alíquota de leite. Este aspecto ilustra o quanto a inclusão de um termo constante pode melhorar a qualidade do ajuste. Não houvesse este termo, a assíntota inferior corresponderia a 0, o que aparentemente não seria uma pressuposição adequada nesta situação.

Também de interesse prático é o valor da temperatura correspondente ao ponto de inflexão. Este valor é estimado como sendo:

$$\frac{-\hat{\beta}_1}{\hat{\beta}_2}=55,2^oC$$

Ou seja, ao se ir aumentando a temperatura, na vizinhança do valor $55,2^oC$ a redução da quantidade de microrganismos acontece com velocidade máxima.

Pode haver interesse também em outros valores de temperatura. Os valores estimados pelo modelo variam de 4,05 a 7,00. Ou seja, com uma diferença de 2,95 entre as duas assíntotas, correspondente à estimativa de β_3. Assim, por exemplo, qual seria a temperatura que deixaria uma população de microrganismos (diga-se) de 20% deste intervalo, acima do mínimo? Em outras palavras, qual o valor de X que teria um valor de Y igual a $4,05+0,2\times 2,95$? Nesta linha de raciocínio, pode-se pensar em qualquer fração p, outras que não 20%. Chamando este valor de Y por Y_c, tem-se:

$$p=\frac{Y_c-\hat{\beta}_0}{\hat{\beta}_3}$$

sendo que p é uma fração desejada, como 0,2. Assim:

$$Y_c=\hat{\beta}_0+p\hat{\beta}_3$$

Igualando Y_c ao modelo ajustado, e colocando X em evidência, tem-se que o valor de X correspondente a Y_c é dado por:

$$X=-\frac{\ln\left(\frac{1}{p}-1\right)+\hat{\beta}_1}{\hat{\beta}_2}$$

Perceba-se que, para $p=0,5$, cai-se na expressão usual do ponto de inflexão da curva.

Assim, para um $p = 0,2$, a temperatura correspondente de interesse é:

$$X = -\frac{\ln\left(\frac{1}{0,2} - 1\right) + 29,5006}{-0,5340} = 57,8^oC$$

o qual, coerentemente, é um valor maior do que aquele obtido anteriormente.

O mesmo método foi realizado para a *Escherichia coli* no processo de ultrassom, tendo-se chegado ao seguinte modelo ajustado:

$$\hat{Y}_i = 0,48 + \frac{6,28}{1 + e^{-(10,06-0,19X_i)}}$$

cujo gráfico é dado na Figura 5.

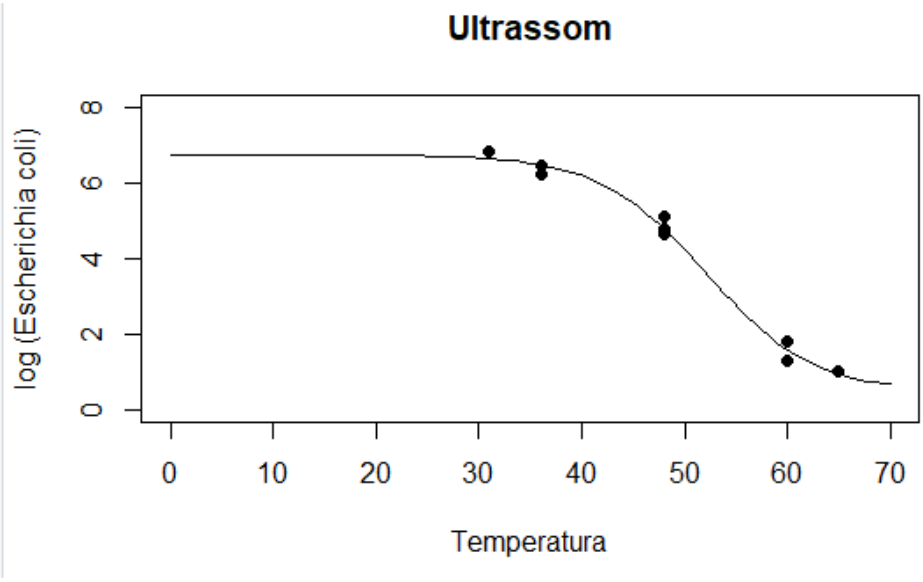

Figura 5: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Escherichia coli* em alíquotas de leite, em função da temperatura de processamento (método: ultrassonificação).

É possível notar, nesse ajuste, que a assíntota superior foi equivalente a 6,76 e a inferior a 0,48. E que a temperatura correspondente ao ponto de inflexão é de $52,94^oC$, correspondendo a uma menor temperatura, no método de ultrassom, quando comparado à pasteurização.

Também chama a atenção o fato de o método ultrassom ter reduzido consideravelmente mais a população da bactéria *Escherichia coli*, em relação à pasteurização, dado o valor bem mais baixo da assíntota inferior.

4.1.2 *Staphylococcus aureus*

A segunda bactéria cujos dados foram ajustados à regressão logística foi a *Staphylococcus aureus*, e seus gráficos, para pasteurização e ultrassom, respectivamente, em relação à temperatura, se encontram nas Figuras 6 e 7. O modelo ajustado para a pasteurização foi:

$$\hat{Y}_i = 4,01 + \frac{2,84}{1 + e^{-(15,64 - 0,27X_i)}}$$

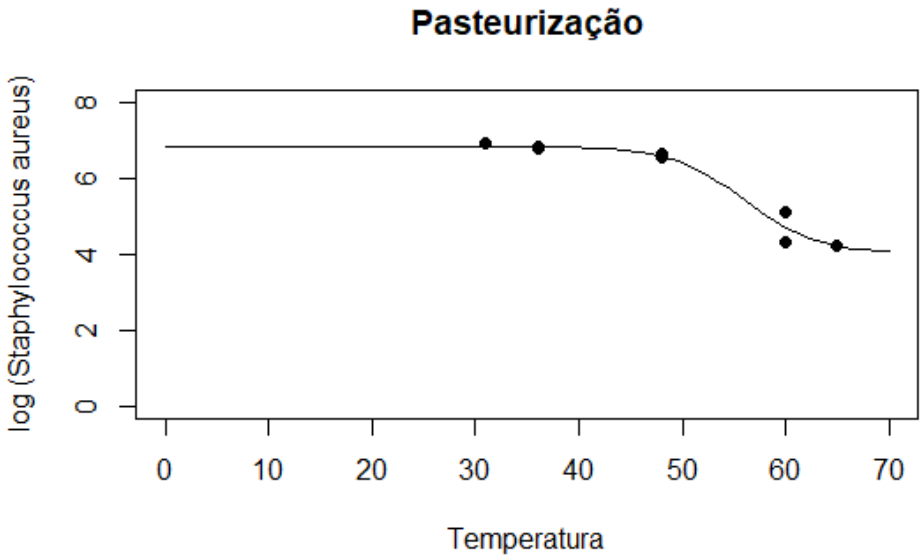

Figura 6: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Staphylococcus aureus* em alíquotas de leite, em função da temperatura de processamento (método: pasteurização).

E para o processo de ultrassom o modelo encontrado foi:

$$\hat{Y}_i = 0,93 + \frac{6,0019}{1 + e^{-(21,64 - 0,40X_i)}}$$

As assíntotas de cada modelo ajustado podem ser apreciadas na Tabela 7.

Tabela 7: Assíntotas referentes ao logaritmo do número de microrganismos *Staphylococcus aureus* do modelo de regressão logística ajustado para dois métodos de processamento.

Método	Assíntota Superior	Assíntota Inferior
Pasteurização	6,85	4,01
Ultrassom	6,99	0,93

Como visto para a bactéria anterior, o leite humano levado ao processo de ultrassom também mostrou-se mais eficiente, no sentido de promover rápidas reduções com menores

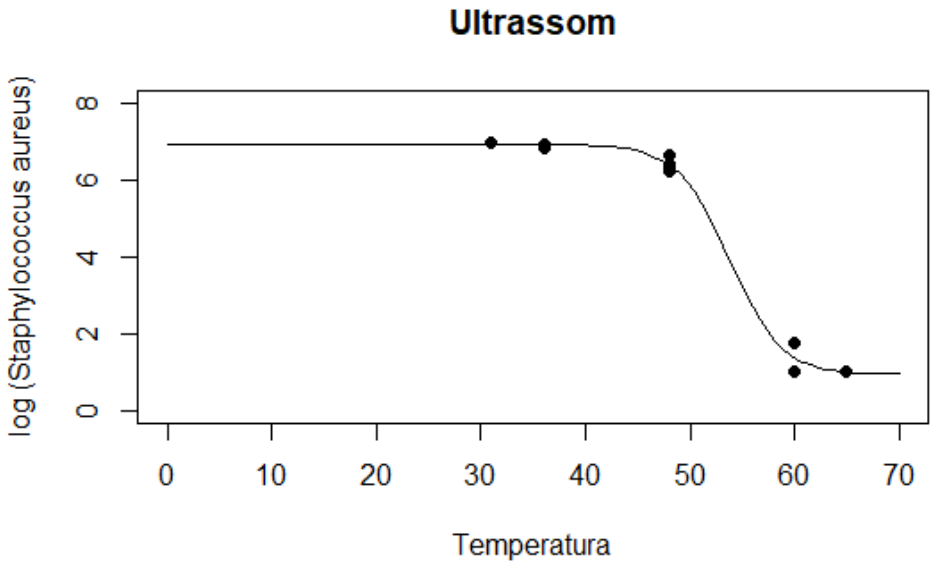

Figura 7: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Staphylococcus aureus* em alíquotas de leite, em função da temperatura de processamento (método: ultrassom).

temperaturas, em relação ao processo térmico. Os pontos de inflexão foram, para a pasteurização e o ultrassom, $57,92^oC$ e $54,1^oC$, respectivamente. Também pode-se verificar na Tabela 7 que o segundo método promoveu uma redução bem maior do microrganismo *Staphylococcus aureus*.

4.1.3 *Salmonella ssp*

O terceiro e último microrganismo considerado corresponderam a bactérias do gênero *Salmonella ssp*. Os gráficos do ajuste à regressão logística, para cada método, estão apresentados nas Figuras 8 e 9. Os modelos ajustados, foram respectivamente:

$$\hat{Y}_i = 3,25 + \frac{3,70}{1 + e^{-(21,62-0,39X_i)}}$$

$$\hat{Y}_i = 1,08 + \frac{7,77}{1 + e^{-(15,05-0,24X_i)}}$$

Os pontos de inflexão para os métodos de pasteurização e ultrassonificação foram iguais a $55,4^oC$ e $62,74^oC$, respectivamente. Ou seja, ao contrário do que aconteceu com os outros dois microrganismos, aqui o ponto de inflexão ocorreu a uma temperatura maior para o método de ultrassonificação. As assíntotas dos modelos ajustados estão apresentadas na Tabela 8. Pode-se perceber que, embora o ponto de inflexão tenha ocorrido em uma maior temperatura, o método de ultrassom propiciou uma maior redução no número

de microrganismos presentes nas alíquotas de leite. Ou seja, pode-se dizer, de uma forma geral, que este método apresentou uma maior eficiência, em relação à pasteurização.

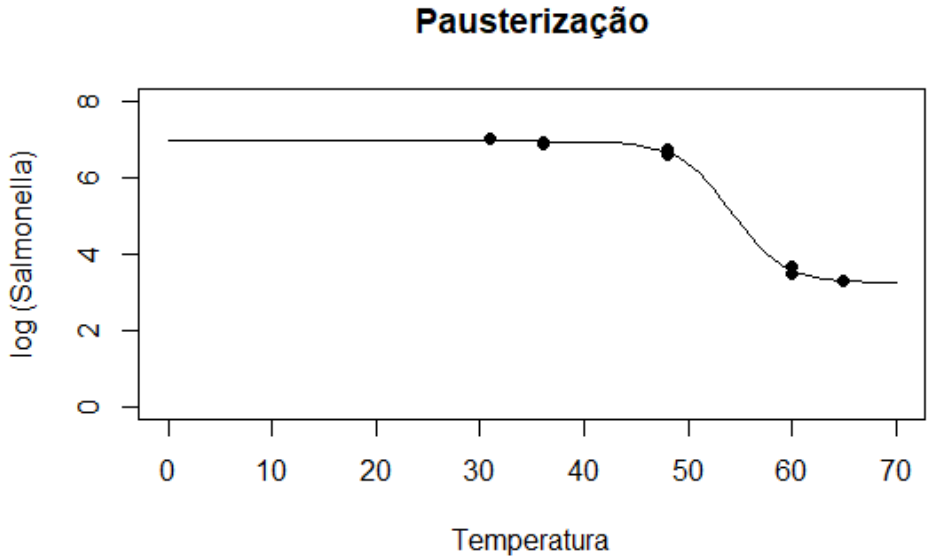

Figura 8: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Salmonella ssp* em alíquotas de leite, em função da temperatura de processamento (método: pasteurização).

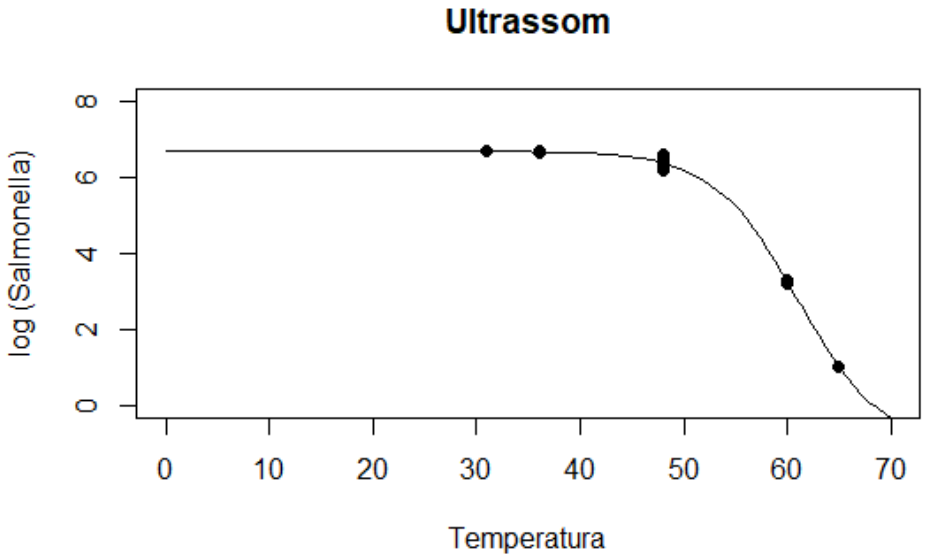

Figura 9: Gráfico do ajuste de um modelo de regressão logística a dados referentes à presença de *Salmonella ssp* em alíquotas de leite, em função da temperatura de processamento (método: ultrassom).

Tabela 8: Assíntotas referentes ao logaritmo do número de microrganismos *Salmonella ssp* em relação à tempetarura

Processo	Assíntota Superior	Assíntota Inferior
Pasteurização	6,95	3,25
Ultrassom	8,85	1,08

4.2 Modelo bisegmentado

Para ilustrar o ajuste de um modelo bisegmentado, utilizou-se a variável resposta: teor de hexanal ($\mu g/mL$) em função do tempo de armazenamento, em dias, presente em alíquotas de leite materno. Foi utilizado um Delineamento Inteiramente Casualizado com três repetições, e o método de processamento utilizado foi a pasteurização. Os dados coletados e disponibilizados por NOGUEIRA (2020) podem ser vistos na tabela 9.

Tabela 9: Dados referentes ao teor de hexanal ($\mu g/mL$) por alíquota de leite, em função do tempo de armazenamento, em dias.

Tempo (dias)	Repetição	Hexanal	Média
0	1	0,1014	
0	2	0,1361	0,1355
0	3	0,169	
15	1	0,7294	
15	2	0,8029	0,7047
15	3	0,5820	
30	1	1,1849	
30	2	1,0501	1,1234
30	3	1,1353	
60	1	0,6328	
60	2	0,7481	0,6718
60	3	0,6345	
120	1	0,4395	
120	2	0,4930	0,4797
120	3	0,5066	

Pode-se observar na Tabela 9 que o teor de hexanal apresentou uma tendência de crescimento até (pelo menos) os 30 dias, sendo que entre os 30 e os 60 dias este teor começou a tender a diminuir. Desta forma, faz sentido pensar em um modelo bisegmentado contínuo (uma vez que o teor de hexanal varia continuamente), em que a intersecção

($X = \theta$) é desconhecida. Iremos admitir, contudo, que θ é algum valor entre 30 e 60 dias.

Inicialmente, tentou-se ajustar um modelo bisegmentado com duas retas, ou seja, o modelo 2.22 com θ desconhecido. Este ajuste não se mostrou adequado (pois os desvios de regressão foram significativos), e assim buscou-se, para o segundo segmento, uma curva (e não uma reta) que explicasse a relação entre X e Y.

Em princípio, poder-se-ia pensar em uma função quadrática para o segundo segmento, uma vez que em ajustes de modelos de regressão linear é comum o uso de polinômios. No entanto, aqui isto teria duas desvantagens. Em primeiro lugar, seriam consumidos todos os graus de liberdade (referentes aos diferentes tempos de armazenamento), para o conjunto de dados em questão. Em segundo lugar, a parábola pode apresentar, na faixa estudada de X, uma relação não monotônica com Y. Em algumas situações isto não é desejado. Por exemplo, poderíamos estar estudando o amadurecimento de frutos de banana, medindo, por exemplo, o teor de clorofila em função do tempo de armazenamento. Espera-se que o teor de clorofila sempre decresça, não fazendo sentido uma curva ajustada (como no caso da parábola com concavidade para cima) que apresentasse uma tendência de crescimento ao final do armazenamento (como se os frutos de banana tendessem a tornar a ficarem verdes, no final da faixa estudada de valores de X, na extremidade superior).

No presente exemplo, após o hexanal ter começado a decrescer, aparentemente não faria sentido este teor tornar a crescer, uma segunda vez. Assim, uma curva mais apropriada, estritamente decrescente (ou estritamente crescente, dependendo do sinal de β_3), seria:

$$Y = \beta_2 + \beta_3 \frac{1}{X} \tag{4.2}$$

Para um valor de β_3 positivo, esta função é estritamente decrescente, diminuindo à medida que X aumenta. A curva desta função apresenta uma assíntota, dada por:

$$\lim_{x \to \infty} Y = \beta_2 \tag{4.3}$$

a qual, no presente exemplo, seria o teor de hexanal residual que tenderia a permanecer no leite, após um longo período de armazenamento. Pode haver interesse particular na estimação deste parâmetro.

Assim, considerando uma reta para o primeiro segmento, e uma curva (dada por 4.2) para o segundo segmento, tem-se que o modelo bisegmentado pode ser expresso conforme o modelo apresentado em 4.4. Por simplicidade de notação, está-se omitindo o termo

residual.

$$Y = \begin{cases} \beta_0 + \beta_1 X, \text{ se } X < \theta \\ \beta_2 + \beta_3 \frac{1}{X}, \text{ se } X > \theta \\ \beta_0 + \beta_1 X = \beta_2 + \beta_3 \frac{1}{X}, \text{ se } X = \theta \end{cases} \tag{4.4}$$

Conforme discutido no Referencial Teórico, também aqui pode-se expressar as duas primeiras equações de 4.4 em um único modelo, fazendo uso de uma variável *dummy*:

$$Z = \begin{cases} 0, \text{ se } X < \theta \\ 1, \text{ se } X > \theta \end{cases} \tag{4.5}$$

e assim as duas primeiras equações de 4.4 podem ser expressas em um único modelo, utilizando o artifício de se subtrair os termos do primeiro segmento dentro do parênteses contendo os termos do segundo segmento:

$$Y = \beta_0 + \beta_1 X + \left(\beta_2 + \beta_3 \frac{1}{X} - \beta_0 - \beta_1 X \right) Z \tag{4.6}$$

de maneira que, para $X < \theta$, tem-se $Y = \beta_0 + \beta_1 X$, e, para $X > \theta$, tem-se $Y = \beta_2 + \beta_3 \frac{1}{X}$. Além disso, devido à continuidade entre os dois segmentos, há uma dependência entre os parâmetros, pois, em $X = \theta$, tem-se:

$$\beta_0 + \beta_1 \theta = \beta_2 + \beta_3 \frac{1}{\theta} \tag{4.7}$$

E assim podemos escrever um parâmetro como função dos demais, por exemplo:

$$\beta_2 = \beta_0 + \beta_1 \theta - \beta_3 \frac{1}{\theta} \tag{4.8}$$

Substituindo 4.8 em 4.6, resulta:

$$Y = \beta_0 + \beta_1 X + \left(\beta_0 + \beta_1 \theta - \beta_3 \frac{1}{\theta} + \beta_3 \frac{1}{X} - \beta_0 - \beta_1 X \right) Z$$

Ou:

$$Y = \beta_0 + \beta_1 X + \left[\beta_1 (\theta - X) - \beta_3 \frac{1}{\theta} + \beta_3 \frac{1}{X} \right] Z \tag{4.9}$$

Tem-se assim um modelo bisegmentado definido em 4.9, mas ainda se faz necessária uma pequena modificação, para o presente exemplo. Neste conjunto de dados, existem resultados para $X = 0$. Como não se pode ter um zero no denominador, pode-se somar uma pequena constante ao denominador dentro do colchete (por exemplo, 0,1), e assim:

$$Y = \beta_0 + \beta_1 X + \left[\beta_1 (\theta - X) - \beta_3 \frac{1}{\theta} + \beta_3 \frac{1}{X + 0,1} \right] Z \tag{4.10}$$

Sendo 4.10, assim, o modelo utilizado no ajustamento. Percebe-se que se trata de um modelo não-linear com 4 parâmetros.

Para a obtenção dos "chutes iniciais", fez-se o ajuste em cada segmento, em separado, obtendo-se estimativas preliminares para os parâmetros, tendo-se ainda partido do valor $X = 45$ dias, como chute inicial do parâmetro θ. As estimativas dos parâmetros foram dadas por (lembrando que $\hat{\beta}_2$ foi obtido como função dos demais):

$$\hat{\beta}_0 = 0,1617 \qquad \hat{\beta}_1 = 0,0327 \qquad \hat{\beta}_2 = 0,2409 \qquad \hat{\beta}_3 = 27,0635 \qquad \hat{\theta} = 30,0000$$

E seu gráfico correspondente pode ser visto na Figura 10.

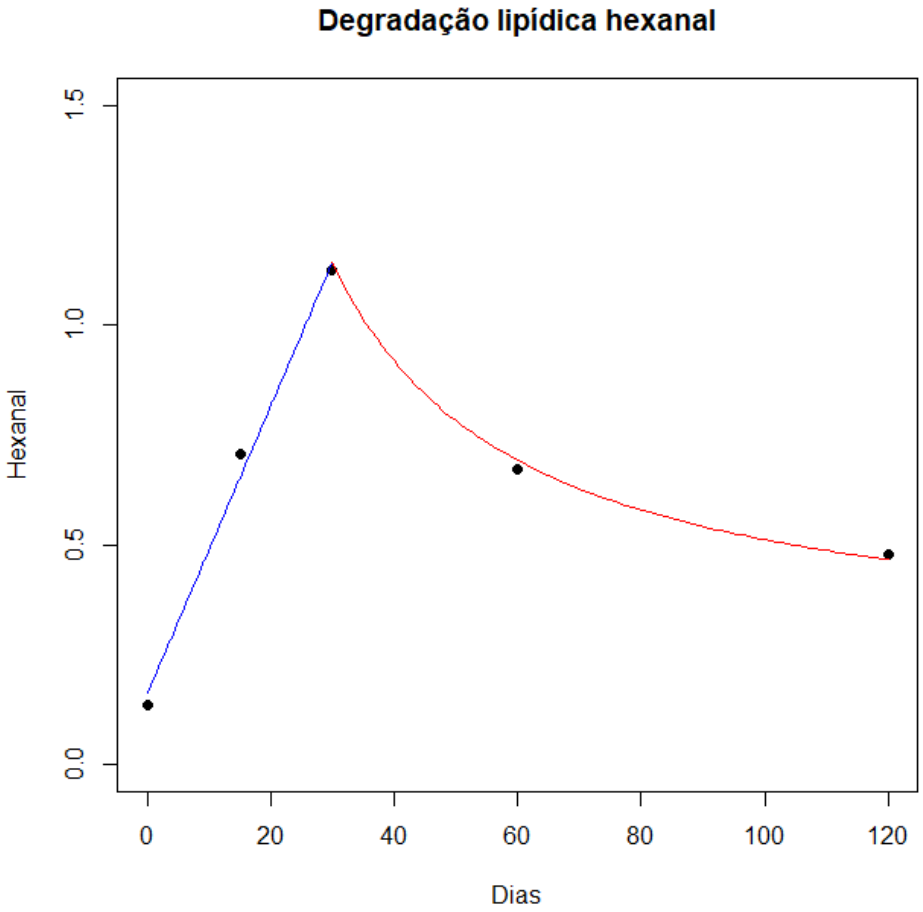

Figura 10: Gráfico do ajuste de um modelo de regressão segmentado contínuo, em dados referentes ao teor de hexanal em alíquotas de leite, em função do tempo de armazenamento, em dias.

Percebe-se, na Figura 10, uma expressiva aderência entre os dados e o modelo proposto. O cálculo do coeficiente de determinação (R^2) e o teste de significância dos desvios de regressão podem ser feitos a partir da ANOVA destes dados, decompondo a variação entre tempos de armazenamento em duas causas de variação: regressão (ajustada para

β_0), e desvios de regressão. Esta ANOVA está apresentada na Tabela 10.

Tabela 10: Análise de variância de dados referentes ao teor de hexanal ($\mu g/mL$) em alíquotas de leite, decompondo a variação entre tempos de processamento em: regressão e desvios de regressão.

Causas de Variação	GL	SQ	QM	F
Tempos	(4)	1,55308	0,33883	80,65***
Regressão (modelo|β_0)	3	1,53979	0,51330	106,72***
Desvios (Falta de Ajustamento)	1	0,01329	0,01329	0,13
Erro Puro	10	0,04814	0,00481	—
Total	14	1,6012	—	—

*** significativo a 0,1% de probabilidade.

Consultando a Tabela 10, verifica-se que o valor do coeficiente de determinação foi elevado:

$$R^2 = \frac{1,53979}{1,55308} = 0,991$$

e que os desvios de regressão (falta de ajustamento) não foram significativos, indicando a adequação do modelo, pois aponta que a variação residual entre tempos (não explicada pela regressão) poderia ser atribuída ao acaso.

5 Considerações finais

Este trabalho procurou realizar uma extensão da disciplina de Regressão, apresentando a análise de regressão não linear, uma possibilidade numérica de ajustamento, exemplos fictícios e reais, destacando tanto aspectos teóricos da análise, como suas possibilidades de aplicação, ilustrando o ajuste com dados oriundos da Escola de Nutrição da UFOP.

Aqui em particular, modelos de regressão não linear foram úteis em duas ocasiões: 1) na elucidação de qual a relação existente entre a temperatura de processamento do leite materno e a proliferação de microrganismos, considerando ainda dois métodos de processamento; 2) a relação entre características lipídicas do leite materno e o seu tempo de armazenamento.

Na primeira situação foram utilizados modelos de regressão logística com resíduos normais, incluindo um termo constante. Os resultados sugerem uma maior eficiência para o método de processamento ultrassonificado. Já na segunda situação, estudou-se uma fração lipídica do leite materno (o teor de hexanal), em função do tempo de estocagem. O ajuste de um modelo bisegmentado não linear propiciou um bom ajuste aos dados, demonstrando um segmento de plena ascensão, seguido de um segmento com decaimento progressivo.

É interessante notar que em ambas as situações foi sempre necessário fazer algumas adaptações, face aos modelos originais, conforme apresentados, por exemplo, em livros-texto. Isto evidencia a necessidade de o estatístico, conhecendo a ferramenta, fazer as devidas modificações, caso necessário.

Dada a grande gama de aplicações da regressão não linear, esperamos ter contribuído para ilustrar sua grande flexibilidade, bem como a necessidade de ampliar sua divulgação.

6 Bibliografia

AGRESTI, A. **Categorical Data Analysis**. 2.ed. New Jersey, Editora John Wiley and Sons, 721p. 2002.

AGUIAR, A.A. **Análise de convergência local do método de Gauss-Newton sob condição Lipschitz**. Goiânia, 2012. 47p. (Monografia de Especialização). Instituto de Matemática e Estatística, Universidade Federal de Goiás.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. John Wiley & Sons, 2014.

HOFFMANN, R. **Análise de Regressão: uma introdução à econometria**. 5.ed. Piracicaba: O Autor, 2016. 393 p.

NOGUEIRA, J.A.V. **Estabilidade lipídica e da atividade antioxidante do leite humano após processamento por termossonicação e armazenamento congelado**. Ouro Preto, ENUT–UFOP, 80p. 2020. (dissertação de mestrado)

PARREIRAS, P.M.; NOGUEIRA, J.A.V.; CUNHA, L.r.; PASSOS, M.C.; GOMES, N.R.; BREGUEZ, G.S.; FALCO, T.S.; BEARZOTI, E.; MENEZES, C.C. Effect of thermosonication on microorganisms, the antioxidant activity and the retinol level of human milk. **Food Control**, v.113, p.107172, 2020.

SILVA, E.M.; FRÜHAUF, A.C.; FERNANDES, F.A.; PAULA, G.S.; MUNIZ, J.A.; FERNANDES, T.J. Método de Newton e Gauss-Newton na estimação dos parâmetros de modelo de regressão não linear. **Sigmae**, v.8, p.728-734, 2019.

APÊNDICE A – Apêndice

Neste Apêndice, são apresentados alguns dos códigos em linguagem **R** utilizados neste material.

A.1 Exemplo: Método de Gauss-Newton

```
# valores x e y:
x<-c(2,4,6,8)
y<- c(4.9,9.3,16.7,24.3)
n<-length(y)

# chutes iniciais
t01 <- 1.8; t02 <- 2
b0 <- c(t01,t02)

# modelo:
mod <- function(x,b1,b2){
  b1+x^b2
}



# Atualizações:
# matriz Z:
Z <- matrix(c(rep(1,n),x^b0[2]*log(x)),nrow = n)

# vetor f:
f <- mod(x,b0[1],b0[2])
```

```
# novo vetor b:
(b1 <- b0 + ginv(t(Z)%*%Z)%*%t(Z)%*%(y-f))

# atualização
b0 <- b1

# volte à matriz Z
```

A.2 Modelo Bisegmentado Descontínuo

```
# 1) modelo segmentado descontinuo
(x <- c(1:8))
(y <- c(1.5,3.1,7.5,10.4,2.4,3.5,5,5.2))

(z <- c(0,0,0,0,1,1,1,1))
(xz <- x*z)

mod3 <- lm(y ~ z+x+xz)
summary(mod3)

# intercepto e inclinacao do primeiro segemento:
mod3$coefficients[1]; mod3$coefficients[3]

# intercepto e inclinacao do segundo segemento:
mod3$coefficients[1]+mod3$coefficients[2]
mod3$coefficients[3]+mod3$coefficients[4]
```

A.3 Regressão Logística (Pasteurização, *Escherichia coli*)

```
# pasteurizacao
dados <- data.frame(Temperatura = c(31, 36, 36, 48, 48, 48, 48, 48, 60, 60, 65),
         y = c(7.04, 6.99, 6.97, 6.96, 6.94, 6.96, 6.88, 6.96, 4.34, 4.2, 4.07))

# chutes iniciais
```

```
auxiliar = log(dados$y/7.04/(1.1-dados$y/7.04));
aux.mod <- lm(auxiliar ~ Temperatura, data = dados)
summary(aux.mod)

library(nls2)
log.mod = nls2(y ~ d + a/(1 + exp(-(b + c*Temperatura))),
               start=list(d=4,a=7.04,b=5.2,c=-0.07),data=dados)
summary(log.mod)

# Somas de Quadrados da ANOVA
# ANOVA entre tempos
mod <- lm(y ~ as.factor(Temperatura), data = dados)
anova(mod)

# Soma de quadrados do Erro Puro
(SQEP <- anova(mod)$Sum[2])

# Soma de Quadrados Total
(SQTC <- var(dados$y)*(length(dados$y)-1))

# Valores preditos pela Regressão Logística:
(yH <- coef(log.mod)[1] + coef(log.mod)[2]/
       (1+exp(-coef(log.mod)[3]-coef(log.mod)[4]*dados$Temperatura)))

# Soma de Quadrados da Regressão, corrigida para a constante
(SQRC <- var(yH)*(length(yH)-1))

# Soma de Quadrados dos Desvios de Regressão
(SQD <- anova(mod)$Sum[1]-SQRC)
```

A.4 Modelo Bisegmentado (Hexanal)

```
# Dados do Hexanal:
(x <- c(0,15,30,60,120))
(y <- c(0.1355,0.7048,1.1234,0.6718,0.4797))
```

```
# obtendo chutes iniciais
(x1 <- x[1:3])
(x2 <- x[4:5])
(y1 <- y[1:3])
(y2 <- y[4:5])

# variavel dummy
z = c(0,0,0,1,1)

mod6 <- lm(y1 ~ x1)
summary(mod6)
mod7 <- lm(y2 ~ I(1/x2))
summary(mod7)

# chute para theta: ponto médio entra 30 e 60
theta=45

library(nls2)
mod8 = nls2(y ~ a + b*x + (b*(theta-x)-d/theta+d/(x+0.1))*z,
+             start=list(a=0.16, b=0.033,d=23.05,theta=45),
+             lower=c(-1,-1,0,30),
+             upper=c(1,1,50,60),algorithm = "port")

summary(mod8)
```