**UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO**

**ANTONIO AUGUSTO CARVAS SANT' ANNA**

# Descrição do intervalo de partos em camundongos de laboratório por meio de modelos lineares generalizados mistos

**CAMPOS DOS GOYTACAZES – RJ**
**MARÇO- 2019**

**ANTONIO AUGUSTO CARVAS SANT' ANNA**

# Descrição do intervalo de partos em camundongos de laboratório por meio de modelos lineares generalizados mistos

Dissertação apresentada ao Centro de Ciências e Tecnologias Agropecuárias da Universidade Estadual do Norte Fluminense Darcy Ribeiro, como requisito parcial para obtenção do grau de Mestre em Ciência Animal na área de concentração de Produção, Reprodução e Saúde Animal.

**ORIENTADOR: LEONARDO SIQUEIRA GLÓRIA**
**COORIENTADOR: RICARDO AUGUSTO MENDONÇA VIEIRA**

**CAMPOS DOS GOYTACAZES – RJ**
**MARÇO-2019**

**Descrição do intervalo de partos em camundongos de laboratório por meio de modelos lineares generalizados mistos**

Dissertação apresentada ao Centro de Ciências e Tecnologias Agropecuárias da Universidade Estadual do Norte Fluminense Darcy Ribeiro, como requisito parcial para obtenção do grau de Mestre em Ciência Animal na área de concentração de Produção, Reprodução e Saúde Animal.

Aprovada em 22 de março de 2019.
Comissão Examinadora:

_____

Dra. Laila Cecília Ramos Bendia (D. Sc., Ciência Animal) - UENF

_____

Dr. Matheus Lima Corrêa Abreu (D. Sc., Ciência Animal) - UFMT

_____

Prof. Moysés Nascimento (D. Sc., Estatística e Experimentação Agropecuária) - UFV

_____

Prof. Ricardo Augusto Mendonça Vieira (D. Sc., Zootecnia) - UENF
(Coorientador)

_____

Prof. Leonardo Siqueira Glória (D. Sc., Genética e Melhoramento) – UENF
(Orientador)

## AGRADECIMENTOS

Gostaria de agradecer a Deus, por permitir trilhar essa caminhada, à minha família que mesmo distante sempre me dá apoio, aos professores Ricardo Augusto e Leonardo Glória por toda paciência e incentivo, à CAPES pelo financiamento do projeto, ao professor Adolpho Marlon e a FIOCRUZ pela parceria.

Minhas amigas Adriana Crispim, Graziele Rodrigues e Bruna Cobuci que sempre me incentivaram, apoiaram e entenderam minha nova etapa.

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de financiamento 001.

Obrigado por tudo!

*“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.”*

*(Arthur Schopenhauer)*

*“Os criadores e os gênios, no início da sua carreira, quase sempre, e muitas vezes até no fim, sempre foram considerados pela sociedade como uns parvos e uns loucos — é esta uma das observações mais triviais e sabidas.”*

*(Fiódor Dostoiévski)*

## RESUMO

**Sant' Anna, Antonio Augusto Carvas, M.Sc., Universidade Estadual do Norte Fluminense – Darcy Ribeiro; Março de 2019; Descrição do intervalo de partos em camundongos de laboratório por meio de modelos lineares generalizados mistos. Professor Orientador: Leonardo Siqueira Glória.**

O objetivo do trabalho foi avaliar a qualidade de ajuste de modelos lineares generalizados mistos para caracterização de animais de laboratório. Os dados dos animais utilizados nesse estudo foram obtidos através da parceria da Universidade Estadual do Norte Fluminense (UENF) com a Fiocruz que são mantidos no biotério da própria Universidade e seguem os procedimentos do comitê de ética da UENF. Analisamos os intervalos de parto, que foram programados no SAS University Edition utilizando a macro %GOF do glimmix. No total foram avaliadas 123 observações, seguindo as distribuições escolhidas, gama, exponencial, normal e log-normal. Foram analisados os gráficos de valores observados versus preditos, gráficos residuais e qualidade de ajuste, a fim de se identificar se o modelo escolhido, juntamente com as distribuições representam a realidade, ou seja, qual é o mais verossímil. A escolha do modelo e das distribuições foram feitas após estudos que levaram em consideração a natureza dos dados e suas características. Os parâmetros da qualidade de ajuste estudados são necessários para selecionar a melhor distribuição para intervalos de partos.



**Palavras – chave:** qualidade de ajuste, glimmix, análise de dados.

**ABSTRACT**

**Sant' Anna, Antonio Augusto Carvas, M.Sc., Universidade Estadual do Norte Fluminense – Darcy Ribeiro; March of 2019; The birth interval of laboratory rats described by a generalized linear mixed effects models. Professor Orientador: Leonardo Siqueira Glória.**

The objective of this study was to evaluate the fit quality of mixed generalized linear models for the characterization of laboratory animals. The data of the animals used in this study were obtained through the partner of the Universidade Estadual do Norte Fluminense (UENF) with Fiocruz that are kept in the laboratory of the University and follow the procedures of the UENF ethics committee. We analyzed birth intervals trait, which were programmed in SAS University Edition using the %GOF of glimmix. In total, 123 observations were evaluated, following the chosen distributions, gamma, exponential, normal and log-normal. After analyzing the graphs of observed values versus predicted, residual graphs and goodness of fit, in order to identify if the chosen model together with the distributions represent reality, ie which is the most likely. The choice of model and distributions was made after studies that took into account the nature of the data and your traits. The parameters of goodness of fit studied, are necessary to select the most likely distribution for birth intervals.

**Keywords :** goodness of fit, Glimmix, data analysis.

# Sumário

## 1. Introdução

O método proposto por Fisher et al. (1923), conhecido como método de análise de variância (ANOVA), apresenta os seguintes limitantes: 1) Independência das observações; 2) Variância constante; 3) Os erros seguem uma distribuição normal. No decorrer do século XX, foi notado que devido a assimetria de alguns dados, outros modelos seriam necessários para explicá-los. Assim, iniciou-se o questionamento da normalidade como sendo ajustada para toda e qualquer situação ou distribuição.

Em Henderson (1953), utilizou três métodos diferentes da ANOVA para estimar a variância, começou a considerar o efeito das variáveis aleatórias, para selecionar o melhor modelo representativo, pois até aquele momento, os efeitos utilizados eram somente os fixos, retirando a limitação dos dois primeiros pressupostos da ANOVA e conceituando os modelos lineares mistos. Sendo a definição de efeitos fixos aqueles fatores que são conhecidos e controlados pelo pesquisador, e todos os efeitos não controláveis, não observados ou não conhecidos considerados como erro ou resíduo, ou seja, efeito aleatório (McCULLOCH et al., 2000). Ao se utilizar efeitos fixos e aleatórios no mesmo modelo, pretendemos evitar a superestimação dos dados.

Com o método da ANOVA sendo questionado, foi preciso avaliar outras distribuições, como por exemplo neste estudo, a família exponencial (gama, log-normal, normal e exponencial) que podem modelar o tempo até a ocorrência de um evento. Com o auxílio dos parâmetros de qualidade de ajuste (gráficos, coeficiente de determinação e coeficiente de correlação), podemos escolher qual o modelo mais verossímil para descrever o intervalo de partos de camundongos da espécie *Mus musculus* produzidos pela FIOCRUZ.

A experimentação animal contribui de maneira expressiva no desenvolvimento da ciência e tecnologia, contribuindo para avanços nas áreas de anatomia, fisiologia, imunologia, entre outros. A utilização de um modelo que prediz as características desses animais, de maneira mais eficiente, tende a favorecer ainda mais os avanços nessas áreas. Diante disso, o objetivo desse trabalho foi realizar um levantamento bibliográfico, utilizar o modelo linear generalizado misto, testar a melhor distribuição para a variável intervalo de parto e determinar assim o modelo mais apropriado.

## 2. Revisão de Literatura

Comumente, em experimentos relacionados a animais, utilizamos resultados ou informações estatísticas para representar pico de produção leiteira, vida reprodutiva de uma espécie, peso dos animais abatidos, tempo que os animais levam até serem abatidos, intervalo de partos e cada dado ou informação tem seu modelo ideal de representação. Modelos de regressão, como os lineares generalizados mistos, são usados para prever ou predizer tais ocorrências e ajudam no momento de planejar ações e/ou evitar um problema.

Todo e qualquer modelo é composto por três componentes, que são: A variável dependente ($Y$), a variável independente ($X$) e a função de ligação, sendo eles definidos como (CORDEIRO, G. M; DEMÉTRIO, C. G. B., 2013):

a) Variável resposta ou variável dependente ($Y_1, Y_2, \dots, Y_j$) = É definida assim que se especificam as medidas a serem utilizadas, podem ser contínuas ou discretas. Sendo contínuas quando os dados podem assumir qualquer valor dentro de um intervalo e discretas quando são números inteiros e positivos, normalmente contagens, número de filhotes ou quantidade de animais abatidos.

b) Variável explicativa ou variável independente ($X_1, X_2, \dots, X_i$) = Participam na forma de soma de seus efeitos, em geral, essas variáveis devem ser não correlacionadas, são os dados;

c) Função de ligação = Relaciona as variáveis dependentes com as independentes. A escolha da função depende do problema a ser estudado e sua natureza, normalmente, cada observação ou dado pode ter uma função de ligação diferente. Devemos também nos atentar na hora da escolha da função de ligação, para que seja compatível com a distribuição proposta para os dados e devemos considerar a facilidade de interpretação do modelo.

Cada termo tem sua função dentro do modelo e sua escolha se deve ao examinar cuidadosamente os dados e analisar suas características, como função de probabilidade, intervalo de variação, natureza contínua ou discreta.

## 2.1 Distribuições de probabilidade

Compreender a distribuição de probabilidade associada ao conjunto de dados a serem analisados é extremamente importante, uma vez que, além de permitir uma caracterização mais precisa dos dados, ainda possibilita um melhor ajustamento dos métodos de análise a serem aplicados.

Define-se variável aleatória aquela representativa dos resultados obtidos num experimento, associada a determinada probabilidade. Considerando o espaço amostral X, determinados pelos eventos $E_1$, $E_2$, ..., $E_n$, com probabilidade $p_1$, $p_2$, ..., n, que estão associados aos números $x_1$, $x_2$, ...$x_n$, em que cada elemento $x_i$ (i=1,2, ..., n) tem correspondência a Ei, define-se o conjunto de valores X de variável aleatória (SCHUSTER & CRUZ, 2013).

As variáveis aleatórias podem ser discretas ou contínuas:

As variáveis aleatórias podem assumir apenas um valor inteiro, incluindo o zero, de maneira que se elas assumirem outros valores não previstos, elas se destroem e perdem a essência de valor. Normalmente, essa variável resulta de contagem, razão pela qual seus valores são expressos por meio de números inteiros não negativos. Ex.: número de filhos (SINDELAR et al., 2014).

Já as variáveis contínuas são úteis para calcular probabilidades referentes ao tempo necessário para se concluir uma tarefa, cujos possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração, podem assumir qualquer valor intermediário entre dois limites de valores inteiros reais. Ex: idade, peso, comprimento, entre outros (SINDELAR et al., 2014). Algumas distribuições relevantes e suas características são apresentadas na Tabela 1.

Tabela 1. Funções de Ligação e distribuições.

| **Distribuição** | **Função de distribuição** | **Espaço paramétrico** | **Média** | **Variância** | **Função de Ligação** | |
|---|---|---|---|---|---|---|
| Normal | $f(y) = \frac{1}{\sqrt{2\pi\sigma}} exp[-(y-\mu)^2/2\sigma^2]$ | $-\infty < \mu < \infty$ $\sigma > 0$ | $\mu$ | $\sigma^2$ | Identidade: | $\eta = \mu$ |
| Poisson | $f(y) = \frac{e^{-\lambda}\lambda^y}{y!} I_{(0,1,\dots)}$ | $0 < p \leq 1$ $(q = 1-p)$ | $\lambda > 0$ | $\lambda$ | Logarítmica: | $\eta = ln(\mu)$ |
| Binomial | $f(y) = \binom{n}{y} p^y q^{n-y} I_{(0,1,\dots,n)}$ | $0 \leq p \leq 1$ $n = 1, 2, 3, \dots$ $(q = 1-p)$ | $Np$ | $npq$ | Logística: | $\eta = ln\frac{\pi}{1-\pi}$ |
| Gama | $f(y) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} I_{(0,1,\dots)}$ | $\lambda > 0$ $r > 0$ | $\frac{r}{\lambda}$ | $\frac{r}{\lambda^2}$ | Recíproca: | $\eta = \frac{1}{\mu}$; $log(\mu)$ |
| Log-normal | $f(y) = \frac{1}{y\sqrt{2\pi\sigma}} exp[-(log_e y - \mu)^2/2\sigma^2] I$ | $-\infty < \mu < \infty$ $\sigma > 0$ | $exp\left[\mu + \left(\frac{1}{2}\right)\sigma^2\right]$ | $\exp[2\mu + 2\sigma^2]$ $-exp[2\mu + 2\sigma^2]$ | Logarítmica: | $\eta = \log(\mu)$ |

Fonte: RESENDE, 2007.

### 2.1.1 Distribuição de Poisson

Essa distribuição indica o número de ocorrências no intervalo de tempo, como não existe limite definido de ocorrências, esta poderá ser de 0,1,2... ∞, logo não negativa.

Consideraremos uma variável aleatória discreta que muitas vezes é útil para calcular o número de ocorrências ao longo de um intervalo de tempo ou espaço específicos. Por exemplo, o número de animais nascidos em cinco anos (SWEENEY et al., 2013).

Os dados devem sempre atender às seguintes propriedades:

(a) A probabilidade de uma ocorrência é a mesma para qualquer intervalo de igual comprimento;

(b) A ocorrência ou não ocorrência em qualquer intervalo é independente da ocorrência ou não ocorrência em outro qualquer intervalo.

Existe uma relação entre a distribuição de Poisson e a exponencial, que pode ser verificada com t (um valor positivo qualquer) para o tempo definido para observação de uma variável aleatória. T são os tempos entre as ocorrência do evento dentro do intervalo t, e X o número de ocorrências do evento no intervalo (0,t). Quando se espera várias ocorrências dentro de t, essas podem se agrupar em um intervalo de tempo menor e as diferenças entre os tempos das ocorrências tendem a uma distribuição gama. A distribuição de Poisson tem também relação com a distribuição normal, e essa relação se estreita quanto maior for λ, onde P(λ) e N(μ,σ) se relacionam por $\mu = \lambda$ e $\sigma = \sqrt{\lambda}$.

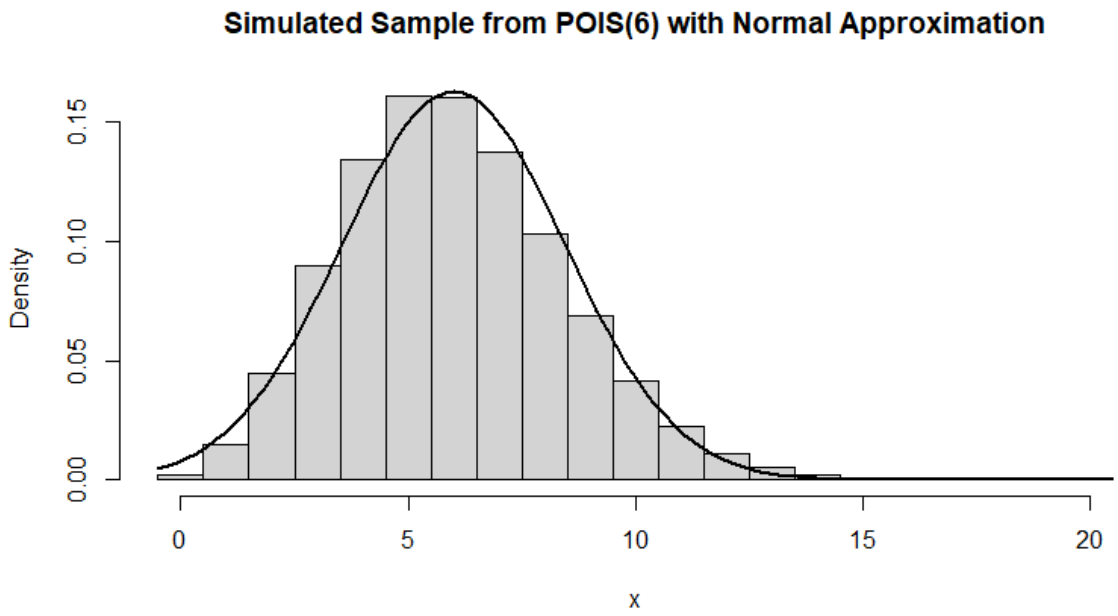


*Figura 1: Distribuição de Poisson.*

### 2.1.2 Distribuição exponencial

Esse modelo é utilizado para variáveis aleatórias, para descrever a extensão do intervalo entre as ocorrências, assume valores não negativos, distribuição assimétrica e mede o tempo entre as ocorrências ou eventos, como exemplo, o tempo para abate, intervalo entre partos, tempo até um equipamento dar defeito. Sua representação gráfica é dada através do gráfico abaixo, em forma de **j** invertido (SWEENEY et al., 2013):

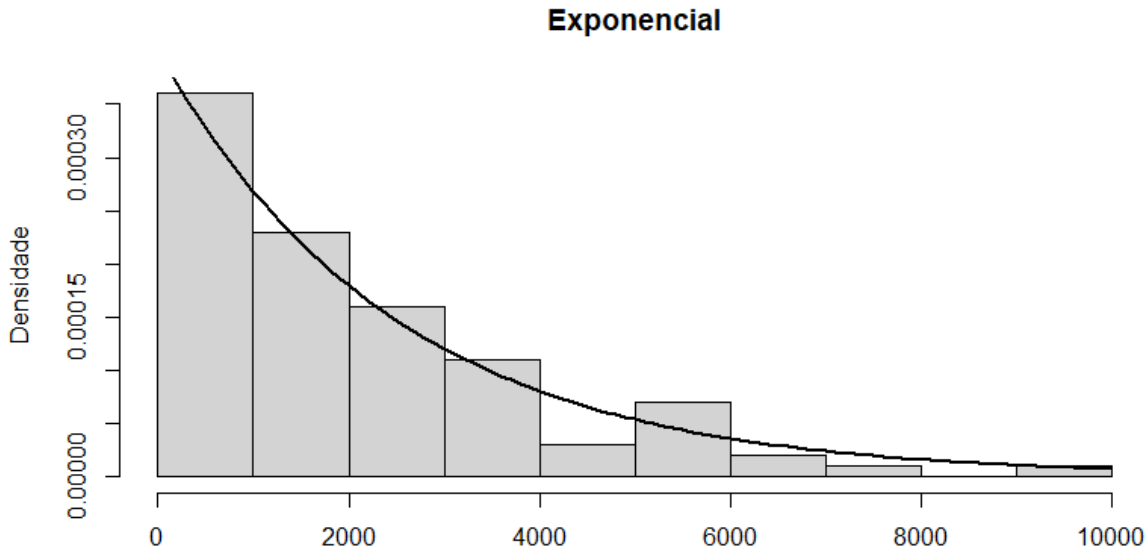


*Figura 2: Gráfico de distribuição exponencial.*

### 2.1.3 Distribuição gama

Cordeiro e Demétrio (2013) descreveram a distribuição gama como sendo comumente utilizada para análise de dados contínuos não negativos que apresentam uma variância crescente e o coeficiente de variação dos dados aproximadamente constante, tais como: tempo de sobrevivência, peso ao abate, tempo até o nascimento, entre outros. Este modelo está associado a dados contínuos assimétricos, com uma cauda exponencial à direita. Pode ser usada para modelar tempos de serviço, vidas de objetos e tempos de reparo. Esta distribuição surge quando indagamos o tempo necessário para obter um número específico de ocorrências do evento.

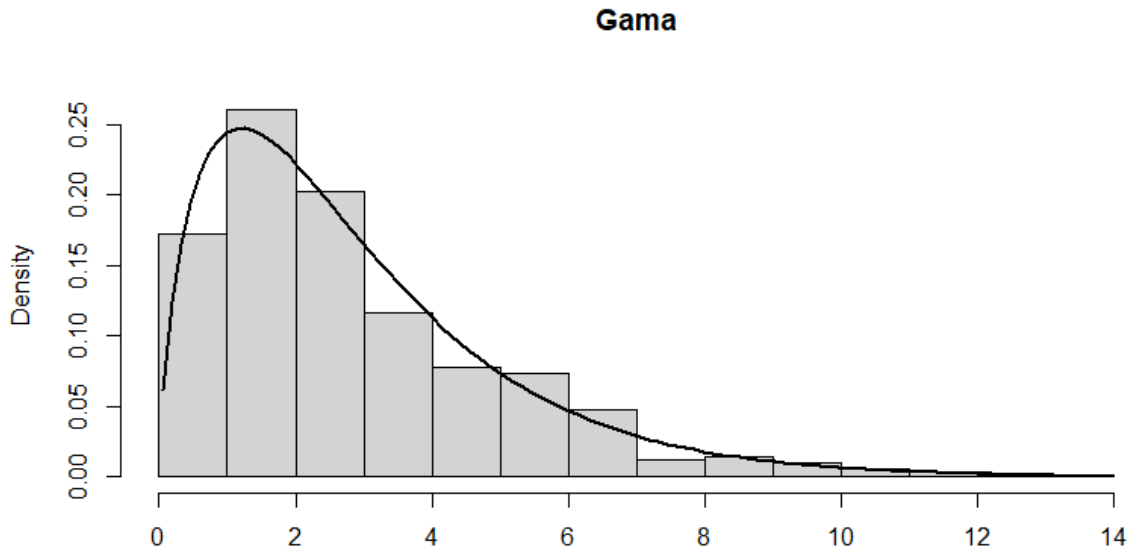


*Figura 3: Gráfico de distribuição gama.*

### 2.1.4 Distribuição normal

A curva normal é definida como sendo simétrica, essa característica encontra-se na natureza quando o número de dados do universo analisado é relativamente grande e principalmente com uma variável contínua. Nesse caso, a distribuição dos valores acontece em uma curva em forma de sino, com um ponto máximo no centro, em que as áreas, em ambos os lados da média, são idênticas.

Essa situação simétrica é estabelecida porque os valores da média, mediana e moda são iguais. Como nem sempre essa situação acontece exatamente dessa forma, comumente usa-se também a expressão de distribuição aproximadamente normal, que se caracteriza por pequenas deformações, em que as medidas da média moda e mediana não são mais iguais, mas com valores muito próximos (SINDELAR et al., 2014).

Em estudo realizado por Studart (2018), a distribuição normal é a mais importante do campo da estatística, uma vez que:

- Serve de parâmetro de comparação;
- Muitas funções convergem para a normal (Poisson, Binomial);
- Muitos fenômenos são descritos pela distribuição normal.

Condições para que uma variável aleatória siga uma distribuição normal (STUDART, 2018).:

- Um grande número de fatores influencia a variável aleatória;
- Cada fator tem, individualmente, um peso muito pequeno;
- Efeito de cada fator é independente dos outros fatores;
- Efeitos dos fatores no resultado é aditivo.

O uso da distribuição normal é devido ao Teorema de Limite Central, que define que "na medida em que o tamanho da amostra aumenta, a distribuição amostral das médias amostrais tende para uma distribuição normal" (TRIOLA, 1999).

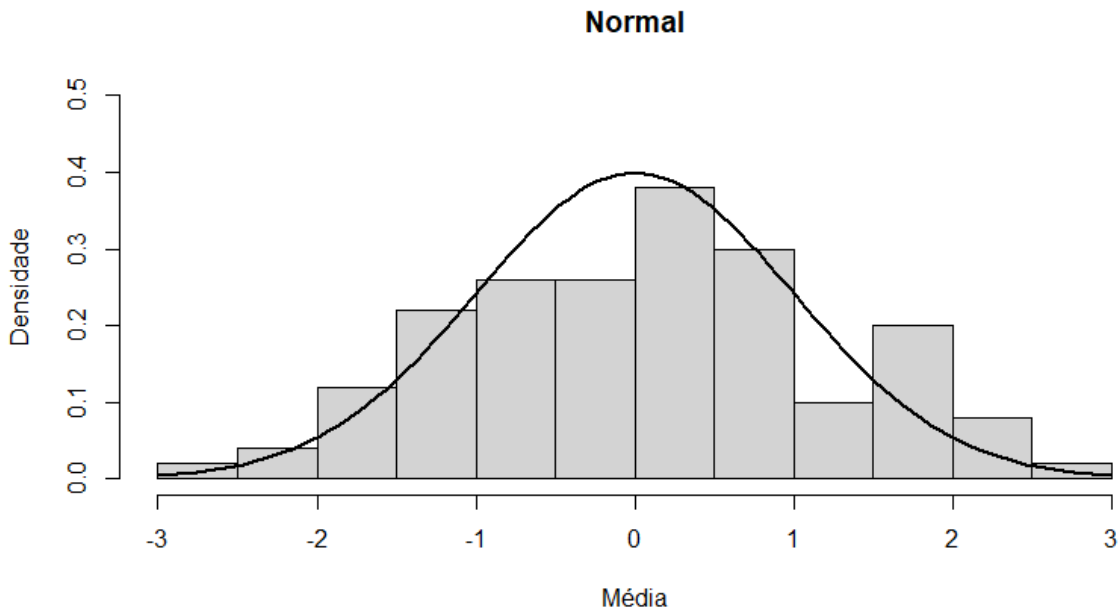


*Figura 4: Gráfico de distribuição normal.*

### 2.1.5 Distribuição log-normal

A distribuição log-normal é assimétrica positiva (deslocada para esquerda) o que a difere da distribuição normal que é simétrica, normalmente é utilizada para modelar o tempo de vida de um produto (validade), o tempo de vida de um objeto até a falha ou fadiga. As distribuições: normal e log-normal são formas de variabilidade, baseadas em forças agindo de forma independente uma da outra, porém, existe uma importante diferença entre elas, seus efeitos podem ser aditivos ou multiplicativos, levando a utilização da distribuição normal ou log-normal respectivamente (MATOS et al., 2010).

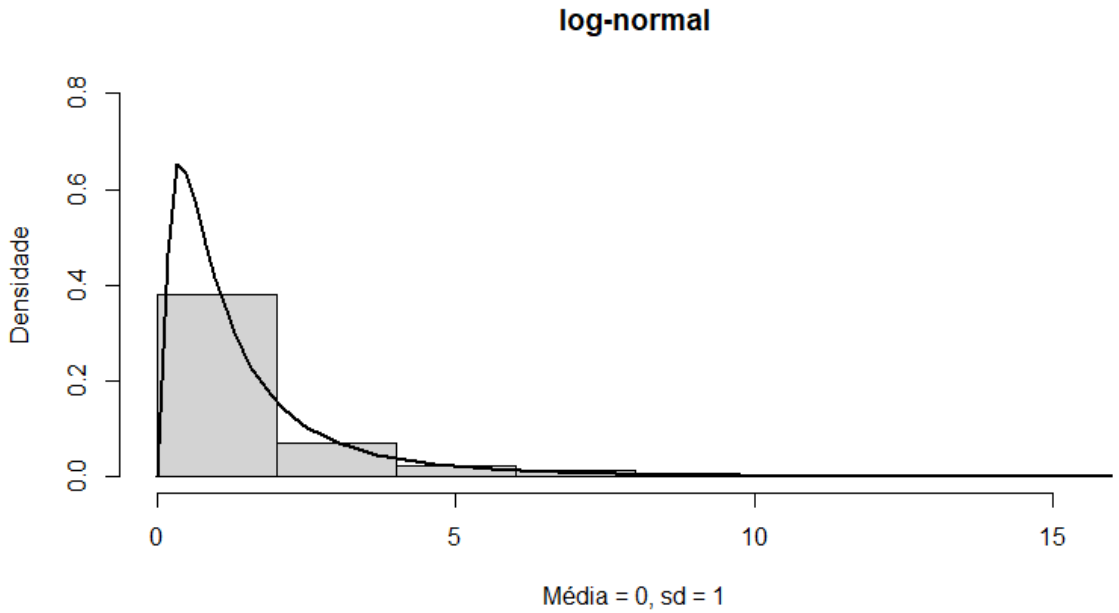


*Figura 5: Gráfico de distribuição log-normal.*

## 2.2 Modelos lineares simples

No modelo linear simples, utilizamos uma variável explicativa responsável por prever ou explicar uma variável dependente, por exemplo, altura e peso, dose de uma droga e resposta, quantidade de adubo e produção de gramíneas (RENCHER, 2000).

Para uma relação linear, usamos um modelo da forma:

$$y = x\beta + e$$

Onde o erro segue uma distribuição normal (N) com média 0 (zero) e variância constante ($\sigma^2$), como representado:

$$e \sim N(0, \sigma^2)$$

No qual y é a variável resposta; $x$ é a matriz de delineamento; $\beta$ é o vetor de parâmetros desconhecidos de efeito fixo e $e$ é o vetor de erro aleatório. Nesse contexto, o erro não significa engano ou equívoco, mas sim um termo estatístico que representa flutuações aleatórias, erros de medidas ou o efeito de fatores não controlados e quando a variável independente não está correlacionada com o resíduo, então ela não é aleatória.

## 2.3 Modelos lineares mistos

O nome modelos mistos vem do fato de que o modelo contém parâmetros de efeito fixo, β, e parâmetros de efeito aleatório, b. Estes modelos são aplicados para modelar a parte aleatória da inclusão de variâncias e covariâncias (LITTELL et al., 2006).

Matricialmente, o modelo linear misto é denominado por:

$$y = x\beta + Zb + e$$

Em que,

$y$ é o vetor de observações;

$x$ é a matriz de incidência dos efeitos fixos (conhecida);

$\beta$ é o vetor de efeitos fixos desconhecidos;

$Z$ é a matriz de incidência dos efeitos aleatórios (conhecida);

$b$ é o vetor de efeitos aleatórios desconhecidos;

$e$ é o vetor de erro aleatórios.

Considera-se que os efeitos aleatórios e os erros (resíduo) têm distribuição normal com média zero e são não correlacionados, com matrizes de variâncias e covariâncias, respectivamente, G e R matrizes positivas definidas, por hipótese, e, portanto, não singulares, dadas por:

$$Var\,(V) = E(vv') = G \quad e \quad Var(e) = E(ee') = R$$

Contudo na prática, é comum nos depararmos com efeitos aleatórios que não seguem uma distribuição normal, nem mesmo com o auxílio de métodos como as transformações. Assim foram propostos os modelos lineares generalizados, que nos permitem trabalhar com outras distribuições.

## 2.4 Modelos lineares generalizados

Nelder & Wedderburn (1972) propuseram uma teoria unificadora da modelagem estatística a que deram o nome de modelos lineares generalizados (MLG), como uma extensão dos modelos lineares clássicos. Na realidade, eles mostraram que uma série de técnicas comumente estudadas separadamente podem ser reunidas sob o nome de Modelos Lineares Generalizados.

A extensão mencionada é feita em duas direções. Por um lado, a distribuição considerada não tem de ser normal, podendo ser qualquer distribuição da família exponencial; por outro lado, embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado e o vector de variáveis pode ser qualquer outra função.

$$y = z\beta = e$$

onde Z é uma matriz de dimensão n × p de especificação do modelo (em geral a matriz de variáveis X com um primeiro vetor unitário), associada a um vetor $\beta = (\beta 1, \ldots, \beta p)^T$ de parâmetros, e $e$ é um vetor de erros aleatórios com distribuição que se supõe $N_n(0, \sigma^2 I)$.

## 2.5 Modelos lineares generalizados mistos (MLGM)

Os Modelos Lineares Generalizados Mistos (MLGM) são uma extensão natural dos Modelos Lineares Mistos (MLM) e dos Modelos Lineares Generalizados (MLG) (McCullagh e Nelder 1989). Os MLGMs, propostos por Breslow e Clayton (1993) são de grande importância, possuem diversas aplicações dada a sua capacidade de modelar a superdispersão dos dados (Williams, 1982) e a dependência entre observações em estudos longitudinais (Stiratelli, Laird e Ware, 1984) ou em dados com medidas repetidas (Breslow, 1984), quando incorporamos efeitos aleatórios.

a. Dado $b_i$ as variáveis respostas $(y_{i1}, \ldots, y_{iTI})$ são mutuamente independentes e seguem um MLG com densidade:

$$f(y_{it}|b_i) = \exp[\frac{\omega_{it}}{\emptyset}\left(y_{it}\theta_{it} - c(\theta_{it})\right) + d(y_{it}, \emptyset)]$$

O valor médio e a variância condicionais são dados, respectivamente, por:

$$E(y_{it}|b_i) = \mu_{it}^b = \frac{\partial_c(\theta_{it})}{\partial\theta_{it}}$$

e

$$E(y_{it}|b_i) = \mu_{it}^b = \frac{\partial_c^2(\theta_{it})}{\partial\theta_{it}^2}\frac{\emptyset}{\omega_{it}}$$

Que se assume que satisfaçam,

$$g(\mu_{it}^b) = x_{it}^b\beta + z_{it}^T b_i$$

e

$$v_{it}^b = V(\mu_{it}^b)\frac{\emptyset}{\omega_{it}}$$

Em que $g$ (.) é uma função de ligação e $V(\mu_{it}^b)$ uma função de variância, ambas conhecidas, $\emptyset$ é um parâmetro de dispersão ou de escala e $\omega_{it}$ é uma constante conhecida;

b. Os efeitos aleatórios, $b_i, i = 1, \ldots, n$ são independentes entre si com uma distribuição multivariada comum *F*.

De modo geral a equação dos MLGM, pode ser escrita na forma:

$$g\{E(y_{it}|b_i)\} = n_{it}^b = x_{it}^T\beta + z_{it}^T b_i$$

com as especificações dadas em 1 e 2, sendo a distribuição comum aos efeitos aleatórios $b_i$ , normal multivariada com valor esperado zero e matriz de covariância *D*.

A inclusão de tais efeitos é uma etapa fundamental na definição do modelo estatístico, pois é por meio do uso de componentes aleatórias que será modelada e inferida a dependência genética existente. Além disto, a classe dos MLGM permite acomodar outras distribuições da família exponencial como as distribuições gama, inversa gaussiana, binomial e poisson, por exemplo, além de permitir o uso de funções de respostas não lineares.

### 2.5.1 Ajuste do Modelo Linear Generalizado Misto

O processo de avaliação dos parâmetros lineares dos modelos, a fase de ajuste é compreendida em três etapas, segundo Cordeiro & Demétrio (2013)**,** que são:

a. Formulação dos modelos:

Nesta etapa devemos examinar e escolher cuidadosamente os dados que iremos utilizar para a distribuição de probabilidade da variável resposta, variáveis explanatórias e função de ligação. Levamos em conta assimetria, natureza contínua, discreta, entre outras características.

De acordo com os dados escolhidos, teremos a melhor função de ligação a ser utilizada, sua melhor aplicabilidade, e logo, o resultado obtido será o mais real, facilitando a interpretação do modelo. Neste momento, ao utilizar dados com menor correlação torna o modelo parcimonioso.

b. Ajuste dos modelos:

Representa o processo de avaliação dos parâmetros lineares dos modelos e de determinadas funções, que representam medidas de adequação dos valores estimados.

Bussab et al., (2017) relatou o seguinte método de ajuste, máxima verossimilhança ou *likelihood* como sendo o mais utilizado. Sendo verossímil tudo que é semelhante à verdade, logo, uma amostra que fornecesse a melhor informação possível sobre um parâmetro de interesse da população, desconhecido, e que desejamos estimar.

c. Inferência:

Consiste em avaliar o modelo escolhido e as discrepâncias existentes. Quando são significativas, podem implicar na escolha de outro modelo, ou em aceitar a existência de observações aberrantes.

Um modelo mal ajustado aos dados, pode apresentar uma ou mais das seguintes condições: (a) inclusão de um grande número de variáveis explanatórias, muitas das quais são correlacionadas e algumas explicando somente uma pequena parcela das observações; (b) formulação de um modelo bastante pobre em variáveis explanatórias, que não revela e nem

reflete as características do modelo; (c) as observações mostram-se insuficientes para que falhas do modelo sejam detectadas.

A condição (a) consiste em uma superparametrização do modelo, (b) é a situação oposta, uma subparametrização que implica em previsões ruins. A terceira condição é um tipo de falha difícil de se detectar, e é devida a combinação inadequada entre distribuição/ função de ligação, que nada tem a ver com as observações em questão.

Os modelos lineares generalizados mistos são bastante práticos, pois na etapa de formulação de modelos têm-se muitas opções de distribuições disponíveis, existem também muitos softwares de fácil utilização e por último, na etapa de inferência, com seus resultados podemos ajustar e retornar nas etapas anteriores, de forma a modificar e trabalhar com modelos mais adequados a necessidade.

## 2.6 Qualidade de ajuste ou goodness of fit (GOF)

Existem várias formas práticas para analisar o modelo mais verossímil, entre elas destacamos a análise gráfica e coeficientes de determinação e correlação, que irão informar a qualidade do ajuste.

### 2.6.1 Análise gráfica

Atualmente, os gráficos são muito utilizados antes e depois que o modelo foi ajustado. A figura abaixo é um exemplo de um gráfico de dispersão que deve ser feito antes de selecionar o modelo (BUSSAB et al., 2017).

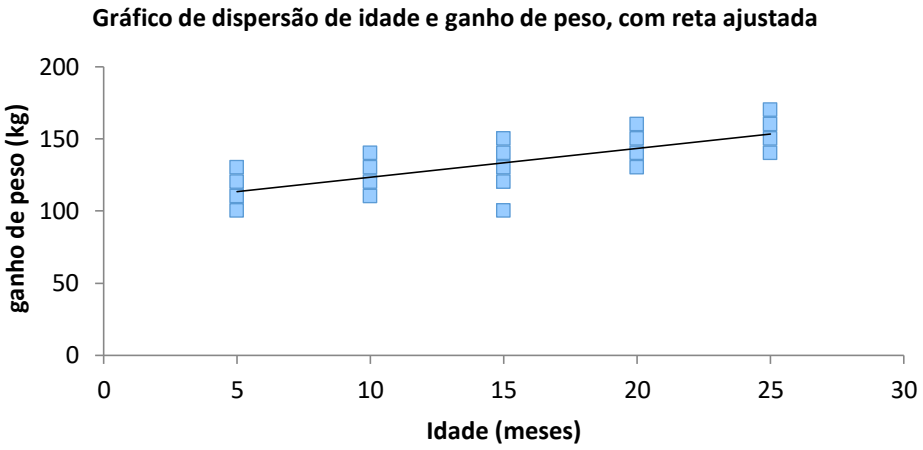


Figura 6: Representação gráfica de ganho de peso e idade.

Esse modelo de gráfico permite visualizar qual a relação entre a variável dependente ($Y$) e a variável independente ($X$), se há valores atípicos (outliers), etc. No caso estudado, utilizaremos duas ou mais variáveis independentes $X_1, X_2, \dots, X_i$, por exemplo, se incluirmos duas variáveis independentes, devemos fazer o gráfico de dispersão entre a variável resposta e cada variável explicativa e entre as duas variáveis $X_1$ *e* $X_2$ (BUSSAB et al., 2017).

A presença de outliers, normalmente se deve ao fato de erro, podendo ser um equipamento mal calibrado, erro de coleta e até mesmo alguma informação ignorada pelo pesquisador. Quando presentes são observações que se encontram nas extremidades, pontos que estão além da dispersão dos resíduos. Ao identificar essa observação aberrante, devemos verificar se é erro ou aquele fato realmente aconteceu sendo um fato isolado, ou seja, a verdadeira causa, pois sua ocorrência pode causar um ajuste enganoso, devendo ser descartado somente se houver evidência de que representa um erro na gravação, um erro de cálculo ou um mau funcionamento de equipamento, pois são situações que não representam a realidade pesquisada (NETER et al., 1974).

O gráfico da variável independente versus resíduo (gráfico residual), nos mostrará o quanto que o nosso modelo escolhido e a distribuição estão representando o que foi observado. Outro importante gráfico a ser considerado são os valores ajustados em relação aos valores observados que são efetivos como indicadores da qualidade do ajuste do modelo (QUININO et al., 2011). Quando o modelo estiver bem ajustado, este gráfico terá uma linha reta e pontos o mais próximo possível (Figura 6).

Ao analisar o preditor residual versus linear, segundo Neter et al. (1974), devemos esperar que os valores das variáveis independentes (X) estejam dispersos horizontalmente e em torno de zero e sem qualquer inclinação, denotando a independência dos erros, ao ocorrer inclinação, a relação pode ser positiva, quando inclinada para direita (Figura 7) ou relação negativa, quando inclinada para esquerda (Figura 8).

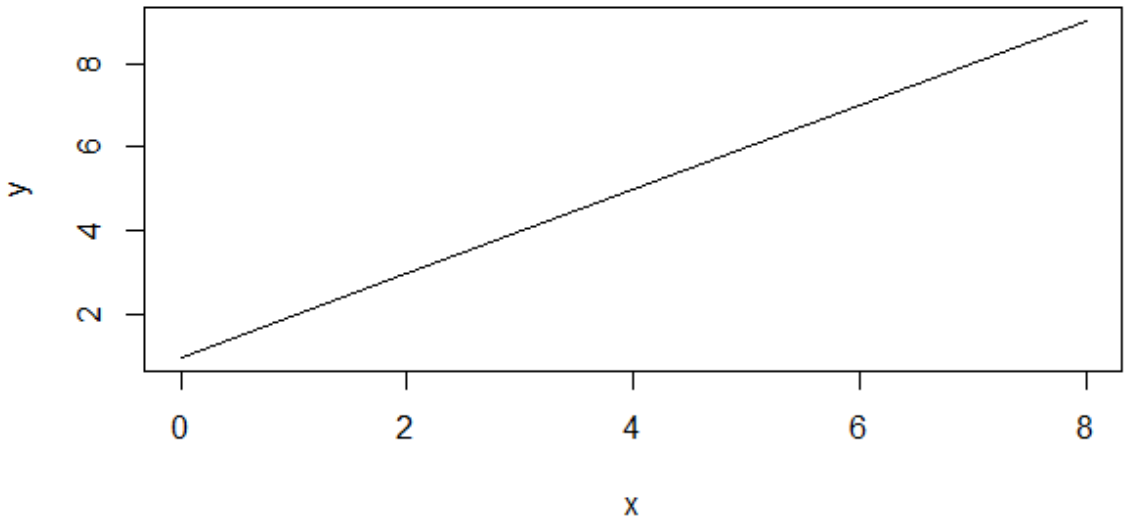


Figura 7. Representação gráfica relação positiva.

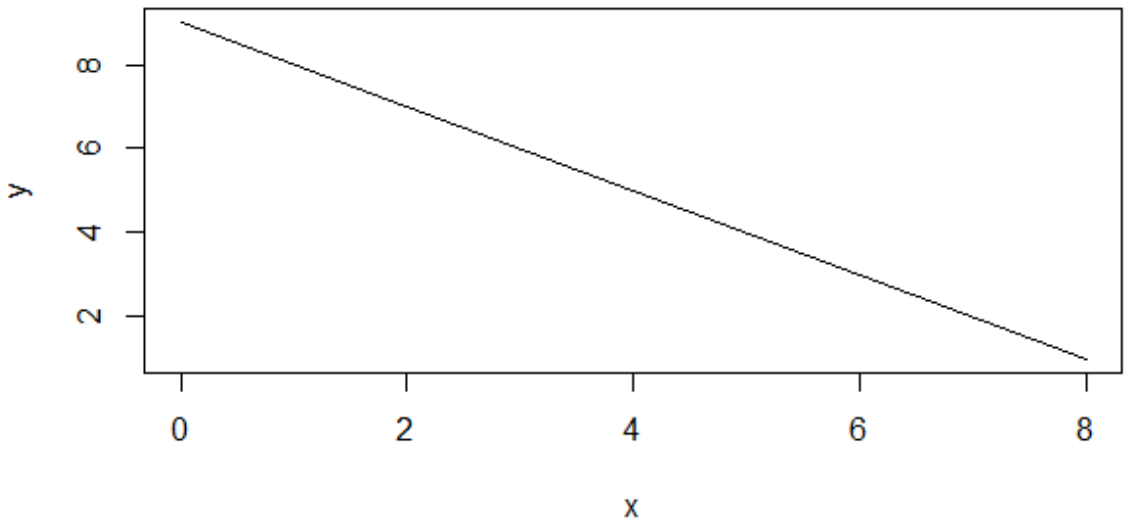


Figura 8. Representação gráfica relação negativa.

### 2.6.2 Coeficiente de Determinação e Correlação

Segundo Vonesh et al. (1996), “estruturas gráficas baseadas em valores observados versus valores preditos oferecem uma alternativa para avaliar a qualidade de ajuste, pois tal gráfico permitirá avaliar visualmente a qualidade de ajuste de um modelo escolhido”. Em seu estudo foi utilizado uma medida de concordância entre as respostas ajustadas e observadas que está ligada ao coeficiente de determinação ($R^2$), ou seja, foi definido a qualidade de ajuste quanto ao grau em que um valor previsto se associa com um valor observado.

O $R^2$ tem seu intervalo de variação entre 0 e 1, é uma medida da qualidade de ajuste e que simplificando temos:

- Ajuste perfeito quando $R^2 = 1$
- Ajuste ruim quando $R^2 = 0$

Também definimos que quanto maior for o valor de $R^2$, maior será a relação entre as variáveis independentes ($X_1, X_2, X_i$) e a variável dependente ($Y$).

O $R^2$ deve ser usado com precaução, pois é sempre possível torná-lo maior pela adição de um número suficiente de termos ao modelo, ou seja, aumentar o número de dados, número de variáveis explicativas, por exemplo. Pode-se também utilizar o $R_a^2$ (coeficiente de determinação ajustado), a fim de evitar uma superestimação, definido como:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

No qual n é o número de observações e p as variáveis independentes.

O $R_a^2$ considera o número de variáveis explicativas que o modelo apresenta. Assim, o $R^2$ tradicional sempre aumentará quando colocarmos mais variáveis explicativas no modelo, em oposição ao $R_a^2$ que pode aumentar ou diminuir e até ser negativo. Descrito na forma de (GUJARATI, 2009):

$$R_a^2 = 1 - k(1 - R^2)$$

Se k > 1, $R_a^2 < R^2$, pode ser negativo.

Pode ser explicado seu valor negativo normalmente devido a dois fatores. Primeiro, o tamanho da amostra é pequeno. Segundo, quando uma nova variável regressora é usada e isso é redundante em relação a uma já presente no modelo (NAKAGAWA et al., 2013).

Existe também outra medida que é conhecida como Coeficiente de Correlação (r), que é muito utilizada para medir o grau de associação entre as variáveis, ou seja, como os X' s e Y se relacionam sua covariância, a semelhança entre as variáveis. Sua representação é simples:

$$r = \pm\sqrt{R^2}$$

Em apenas duas situações o r poderá ser igual a zero, quando não tiver relação entre as variáveis ou quando não for linear (GUIMARÃES, 2018).

- r = 1, quando a relação é positiva e perfeita;
- r = -1, quando a relação é negativa e perfeita;
- r = 0, quando não há relação ou a correlação é não linear.

De acordo com Vonesh et al (1996), existem vantagens ao usar o coeficiente de correlação de concordância (r):

1) r é interpretável diretamente como um coeficiente de correlação de concordância entre valores observados e previstos.

2) Os valores possíveis de r estão no intervalo -1< r <1 com um ajuste perfeito correspondente a um valor de um e uma falta de ajuste correspondente a valores menores que 0 (zero).

Pode-se calcular também o $R^2$condicional que leva em consideração a variância que os efeitos fixos e aleatórios designam, por sua vez, $R^2$descreve somente a variância explicada pelos efeitos fixos.

É importante avaliarmos os valores condicionais no momento de realizar a qualidade de ajuste, pois assim o ajuste fica associado aos valores condicionados. Assim como $R^2$o coeficiente de correlação (r) irá aumentar com modelos mais completos ou com mais dados, indicando que necessitamos ajustar seu valor, levando em consideração o número de parâmetros. O r do modelo é utilizado para identificar quais são os efeitos fixos apropriados na avaliação da qualidade de ajuste (VONESH et al., 1996).

## 3. Material e métodos

Foram realizadas 123 avaliações de animais da espécie *Mus musculus*, dados históricos do biotério da Fiocruz. Avaliamos a característica de intervalos de partos, usamos a teoria de modelos lineares generalizados mistos pelo procedimento GLIMMIX do software SAS e a escolha da função de distribuição estatística será pela macro% GOF do mesmo software.

A macro GOF utilizada pelo SAS, nos auxilia a gerar dados como o coeficiente de determinação ($R^2$) e o coeficiente de correlação ($r$), ambos apresentam variações condicionais e ajustadas. Primeiramente, $R^2$ nos informa o quanto que a variável $Y$ é explicada pelas variáveis independentes de $X's$. O coeficiente de correlação, nos informa como $X$ e $Y$ se relacionam, é uma medida de associação entre variáveis (VONESH et al., 1996).

O $R^2$ condicional, em seu cálculo, leva em consideração efeitos fixos e aleatórios, sendo, portanto, muitas vezes menores que o valor de $R^2$ tradicional, que considera apenas os efeitos fixos. Já os valores ajustados consideram o número de parâmetros e o número de observações, para evitar superestimação.

Nakagawa et al., (2013), explicou de uma forma muito simples a utilização de $R^2$, "por não possuir uma unidade, pode-se usá-la para avaliar o ajuste de modelos e comparar os valores de $R^2$em todos os estudos".

$$R^2 = 1 - \frac{\sum (\mathrm{yi} - \hat{\mathrm{y}}\mathrm{j})2}{\sum (\mathrm{yi} - \bar{\mathrm{y}})2} \qquad \text{( eq. 1 )}$$

$$R_a^2 = 1 - k\,(1 - R^2) \qquad k = \frac{N}{N-S} \qquad \text{( eq. 2)}$$

N = número de observações

S = número de parâmetros

$$R_c^2 = 1 - \frac{\sum (\mathrm{yi} - \hat{\mathrm{y}}\mathrm{j})2}{\sum (\mathrm{yi} - \bar{\mathrm{y}})2} \qquad \text{( eq. 3 )}$$

$$R_{c,a}^2 = 1 - k\,(1 - R_c^2) \qquad \text{( eq. 4 )}$$

$$r = 2\,\frac{R^2}{1+R^2} \qquad \text{( eq. 5 )}$$

$$r_a = 1 - k\,(\,1 - r\,) \qquad \text{( eq. 6 )}$$

$$r_c = 2\,\frac{r_c^2}{1+r_c^2} \qquad \text{( eq. 7 )}$$

$$r_{c,a} = 1 - k\,(\,1 - r_c\,) \qquad \text{( eq. 8 )}$$

## 4. Resultados e discussões

Modelos lineares generalizados são bastante práticos, já que no estágio de formulação de modelos existem muitas opções de distribuição disponíveis, existem também muitos softwares fáceis de usar e por último, na etapa de inferência, com seus resultados podemos ajustar e retornar nas etapas anteriores a fim de modificar e trabalhar com modelos mais adequados à necessidade, caso o modelo inicial não seja adequado.

Ao analisar os dados, segundo Bolker et al., (2009), devemos sempre seguir alguns passos: escolher uma distribuição, analisar os gráficos e as respostas. Para este trabalho, escolhemos as distribuições exponencial, gama, normal e log-normal.

### 4.1 Análise gráfica

Observamos que os gráficos residuais (figura 10) das distribuições lognormal, gama e exponencial, os dados se dispersam na horizontal e em torno de zero, como esperado, porém não foi possível compará-los, pois são retratados em diferentes escalas. Se houver alguma inclinação, indicará correlação positiva ou negativa e não esperamos esse fato, mas sim, que os erros são independentes e o modelo linear utilizado é apropriado.

Outro dado importante, nenhum dos gráficos apresentou *outliers*, o que sugere que os dados utilizados não sofreram erro de medição, coleta ou mau funcionamento do equipamento.

O gráfico da distribuição gaussiana, tem uma inclinação, que indica que o erro é maior quando aumenta o valor de X, que denota uma dependência residual e que a variação do erro não é constante.

Ao analisar a figura 9, podemos observar que a distribuição normal foi a que apresentou valores preditos o mais próximo de valores observados, como demonstrado pela proximidade da reta de unidade. Mesmo não esperando que tal distribuição fosse a indicada para representar intervalo de partos, pois é uma característica de uma distribuição exponencial, por exemplo, o gráfico nos indica que a distribuição normal ainda é a mais indicada.

## 4.2 Qualidade de ajuste (GOF)

A maioria dos pesquisadores deseja ter um modelo o mais adequado possível, o que significa que o modelo apresentado serve para descrever tais características e prever situações futuras de acordo com cada realidade ou situação (NETER et al., 1974). Há muita discussão sobre dois coeficientes que estão sendo amplamente estudados: coeficiente de determinação ($R^2$) e coeficiente de correlação ($r$).

A figura 11 mostra o output do SAS, suas respectivas distribuições, os coeficientes e suas variações, que são marginais e condicionais. Os modelos marginais, estudados por Kuruso (2013), descrevem dados populacionais e levam em conta apenas os efeitos fixos. Os modelos condicionais, no entanto, descrevem o indivíduo, considerando também os efeitos aleatórios.

Ao analisar os dados do GOF após a execução no SAS, preferimos usar os dados condicionais, porque eles são mais precisos, mais plausíveis e menos distorcidos. Pode-se ver que a distribuição normal apresentou os maiores e melhores resultados de $R^2$, $R_a^2$, e $r_a$, todos próximos de 1. No entanto, os valores de GOF não ajudaram na escolha da melhor distribuição a ser utilizada, e mais ajustes foram necessários para desenvolver o melhor modelo.

## 5. Conclusão

Os parâmetros de adequação do ajuste estudados são necessários para selecionar a distribuição mais provável para os intervalos de partos em camundongos.

## 6. Referências Bibliográficas

BOLKER, B. M; BROOKS, M. E; CLARK, C. J; GEANGE, S. W; POULSEN, J. R; STEVENS, M. H. H; WHITE, J. S. S. **Generalized mixed models: A practical guide for ecology and evolution.** Trends in Ecology & Evolution USA: 2009.

BRESLOW, N. E. "Extra-Poisson variation in log-linear models". Em: Applied Statistics, pp. 38–44. 1984.

BRESLOW, N. E E CLAYTON D. G. "**Approximate inference in generalized linear mixed models**". Em: Journal of the American Statistical Association 88.421, pp. 9–25. 1993.

BUSSAB, W. O; MORETTIN, P. A. **Estatística básica.** 9 ed. São Paulo: Saraiva, 2017.

CORDEIRO, G. M; DEMÉTRIO, C. G. B. **Modelos lineares generalizados e extensões.** Piracicaba: 2013.

CHARLES E. McCULLOCH; SHAYLE R. SEARLE; JOHN M. NEUHAUS. **Generalized, Linear and Mixed models.** Wiley, 2000.

FISHER, R. A.; MACKENZIE, W. A. **Studies in crop veriation II. The manurial response of different potato varieties**. Journal of Agricultural Science, Cambridge, v. 13, p. 311-320, 1923.

GUIMARÃES, P. R. B. **Análise de correlação e medidas de associação.** Prof UFPR material retirado do site http://docs.ufpr.br/~jomarc/correlacao.pdf Acessado: 02.03.2018

GUJARATI, D. N. **Basic Endometrics**. 4 ed. New Delhi: 2009.

HENDERSON, C. R. **Estimation of variance and covariance components**. Biometrics, 1953.

LITTELL, R.; MILLIKEN, G.; STROUP, W.; WOLFINGER, R.; SCHABENBERGER. O. **SAS for mixed models**. New York, 6 ed. Cary, N.C.: SAS Institute, 2006.

MATOS, P. Z; ZOTTI, D. M. **Análise de confiabilidade aplicada à indústria para estimações de falhas e provisionamento de custos.** Curitiba: 2010. Monografia apresentada à disciplina de Laboratório de estatística do curso de estatística do setor de ciências exatas da Universidade Federal do Paraná.

MCCULLAGH, P. AND NELDER, J.A. **Generalized Linear Models**. 2nd edition, Chapman and Hall, London. 1989.

NAKAGAWA, S; SCHIELZETH, H. **A general and simple method for obtaining $R^2$ from generalized mixed-effects models**. Methods in Ecology and Evolution. 2013.

NELDER, J. A; WEDDERBURN, R. W. M. **Generalized linear models.** Journal of the Royal Statistical Sciety, 1972.

NETER, J; WASSERMAN, W. **Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs**. University of Georgia: 1974.

QUININO, R. C; REIS, E. A; BESSEGATO, L. F. **O Coeficiente de Determinação $R^2$ como Instrumento Didático para Avaliar a Utilidade de um Modelo de Regressão Linear Múltipla**. Brasil: 2011.

RENCHER A. C. **Linear Models in Statistics.** New York: Willy International Science, 2000.

RESENDE, M. D. V. **Matemática e estatística na análise de experimentos e no melhoramento genético**.1edição. 362p. 2007.

SCHUSTER, I.; CRUZ, C.D**. Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. 2edição, 1reimpressão. Editora UFV. 2013.

SINDELAR, C. W; CONTO, S. M; AHLERT, L. **Teoria e prática em estatística para cursos de graduação.** 1 ed. Lajeado: Univates, 2014.

STIRATELLI, E.; LAIRD, N.; WARE, J. H. (1984). **Random-Effects modela for serial observations with binary response.** Biométrica 40, 961-911.

STUDART, T. M. C. **Distribuições de probabilidades contínuas.** Prof UFC, **http://www.cearidus.ufc.br/Arquivos/Prob%20e%20Estat%EDstica/Apostila/Cap%EDtulo%207_Dist%20Cont_completo.pdf** acessado 01.03.2018.

SWEENEY, D. J; WILLIAMS, T. A; ANDERSON, D. R. **Estatística aplicada à administração e economia.** 3 ed. Brasil: Trilha, 2013.

TRIOLA, M. F.(1999). **Introdução à estatística.** 7 Ed. Rio de janeiro: JC editora.

VONESH, E. F. **Goodness-of-Fit in Generalized Nonlinear Mixed-Effects Models.** International Biometric Society. 1996.

WILLIAMS, D. A. "**Extra-binomial variation in logistic linear models**". Em: Applied statistics, pp. 144–148. 1982.

Apêndice

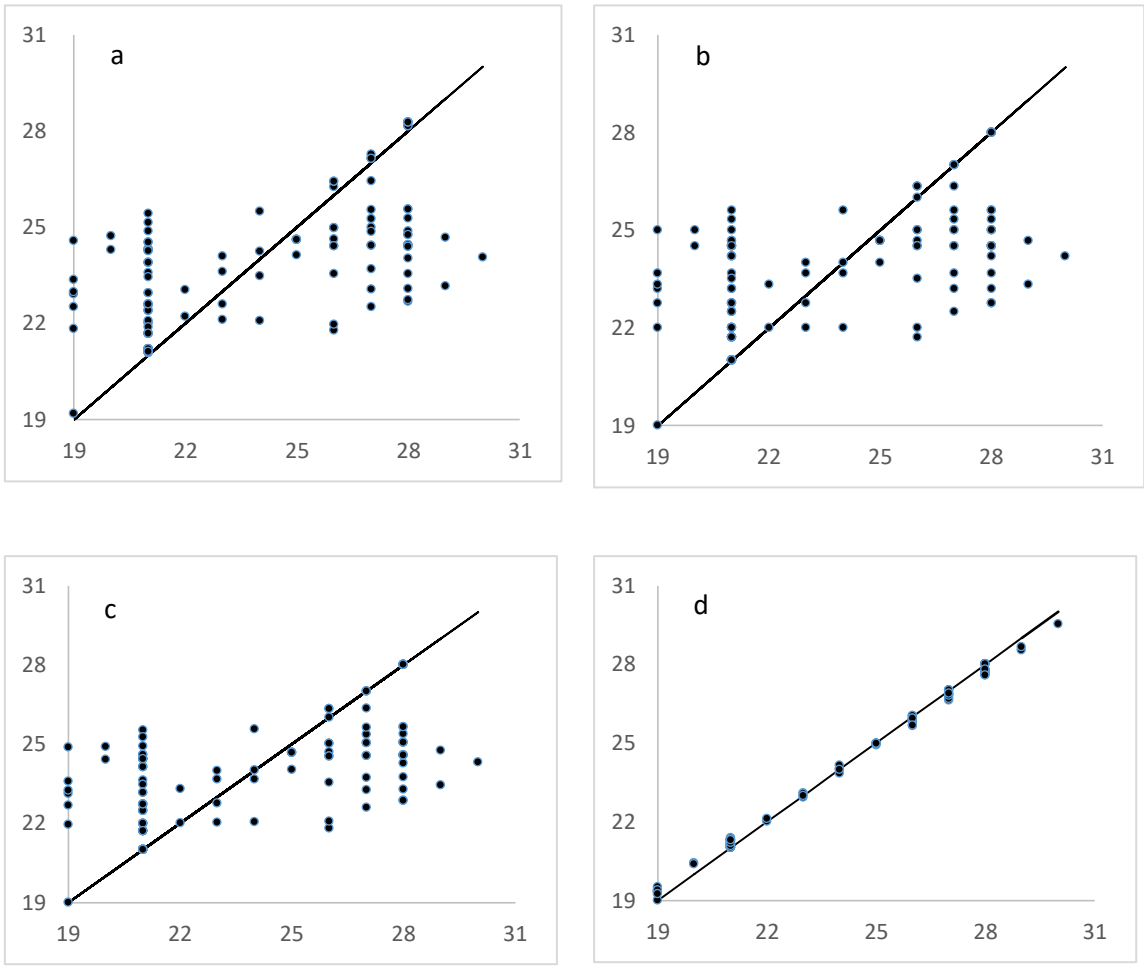


**Figura 9. Gráficos de valores preditos versus valores observados. Distribuição log-normal (a); Distribuição exponencial (b); Distribuição gama (c); Distribuição normal (d).**

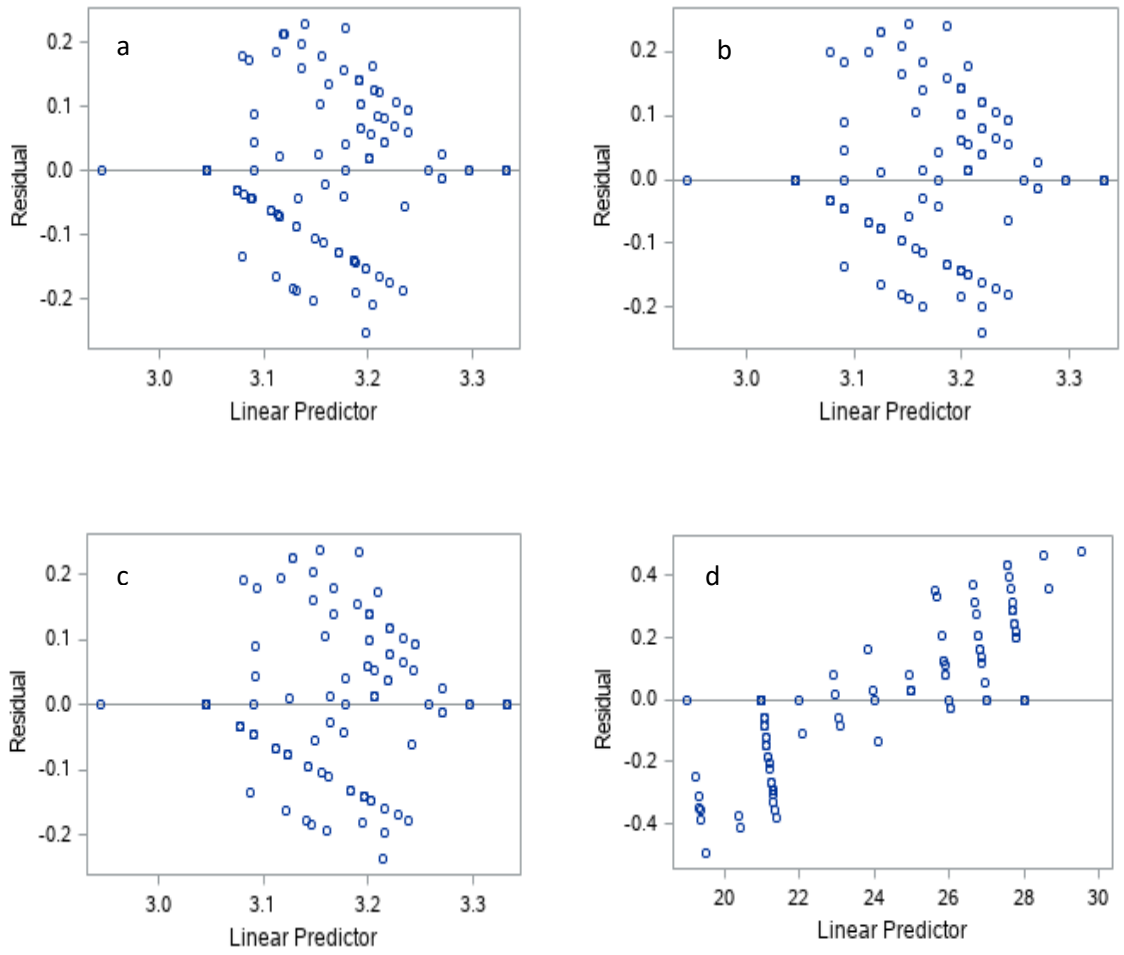


**Figura 10. Gráfico de resíduo. Distribuição log-normal (a); Distribuição exponencial (b); Distribuição gama (c); Distribuição normal (d).**

| Descrição | Exponencial | Gama | Normal | Log-normal |
|---|---|---|---|---|
| Total Observations | 123 | 123 | 123 | 123 |
| N (number of subjects) | 43 | 43 | 43 | 43 |
| Number of Fixed-Effects Parameters | 48 | 48 | 48 | 48 |
| Average Model R-Square | 0.342216 | 0.342212 | 0.342216 | 0.339546 |
| Average Model Adjusted R-Square | -0.07877 | -0.07877 | -0.07877 | -0.083143 |
| Average Model Concordance Correlation | 0.509927 | 0.50989 | 0.509927 | 0.506957 |
| Average Model Adjusted Concordance Correlation | 0.19628 | 0.19622 | 0.19628 | 0.191410 |
| Conditional Model R-Square | 0.342216 | 0.367199 | 0.995528 | 0.364689 |
| Conditional Model Adjusted R-Square | -0.07877 | -0.03779 | 0.992665 | -0.041908 |
| Conditional Model Concordance Correlation | 0.509927 | 0.528593 | 0.997641 | 0.534465 |
| Conditional Model Adjusted Concordance Correlation | 0.19628 | 0.226893 | 0.996131 | 0.236523 |

**Figura 11. Gof**