UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO
CENTRO DE CIÊNCIAS E TECNOLOGIAS AGROPECUÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA ANIMAL

Modelos lineares generalizados mistos para caracterização de animais de laboratório

Mestrando: Antonio Augusto Carvas Sant' Anna
Orientador: Prof. Leonardo Siqueira Glória

CAMPOS DOS GOYTACAZES – RJ
MARÇO- 2018

Modelos lineares generalizados mistos para caracterização de animais de laboratório

PROJETO DE DISSERTAÇÃO

CAMPOS DOS GOYTACAZES – RJ

MARÇO- 2018

SUMÁRIO

1. Introdução 5
2. Objetivo.....6
3. Revisão de literatura.....6
   3.1 Modelos lineares simples.....7
   3.2 Modelos lineares generalizados mistos.....8
      3.2.1 Distribuição de Poisson.....9
      3.2.2 Distribuição de probabilidade exponencial.....10
      3.2.3 Distribuição de probabilidade Gama.....11
      3.2.4 Distribuição normal.....12
      3.2.5 Distribuição de probabilidade log-normal.....14
4. Ajuste do MLG.....14
   4.1 Qualidade de ajuste ou *goodness of it*.....15
5. Materiais e métodos.....18
6. Referências.....19
7. Cronograma.....20

LISTA DE FIGURAS

Figura 1: Distribuição de Poisson e distribuição exponencial. 10
Figura 2: Gráfico de distribuição exponencial 10
Figura 3: Gráfico de distribuição exponencial 11
Figura 4: Gráfico de distribuição gama. 12
Figura 5: Gráfico de distribuição normal. 13
Figura 6: Gráfico de distribuição normal. 13
Figura 7: Representação gráfica de ganho de peso e idade. 16

1. Introdução

A estatística é uma ciência que nos permite analisar determinados eventos, fatos, dados e os quantifica, através de resultados, tornando-os capazes de nos informar qual será a melhor medida tomar, ou seja, é a transformação de dados em informação que permite decisões mais seguras.

“Apesar de existirem indícios sobre a realização de censos na Babilônia, China e Egito desde 3000 anos A.C., apenas no século XVII a estatística passou a ser considerada uma disciplina autônoma.” (SINDELAR 2014).

No mesmo estudo, Sindelar também definiu que atualmente, devido a necessidade de velocidade nas informações, busca por tecnologias, estudo de mercado, as empresas investem cada vez mais para otimizar o processo produtivo.

Com os dados transformados em informações, podemos controlar, por exemplo, a qualidade do produto, mercado consumidor, aspectos da vida econômica, condições climáticas... mostrando que a estatística atua em diversas áreas.

Temos utilização dos modelos de distribuições nos mais diversos setores, seja na economia, produção das fábricas, bolsa de valores, mercado consumidor, na vida política, então, “a estatística está presente em todas as ciências que se envolvem com coleta e análise de dados...” (SINDELAR, 2014).

Os animais de laboratório vêm sendo usados na pesquisa biomédica por mais de 500 anos. São responsáveis pelos avanços nas áreas de anatomia, fisiologia, imunologia e etc. Com essa evolução nos estudos, os pesquisadores passaram a exigir que esses animais tenham condições ideais que atendam as qualidades genéticas, sanitários e mantidos em ambiente controlado, para que possam garantir a confiabilidade do experimento (ANDRADE, 2002).

Os animais de laboratório, segundo Andrade (2002), devem apresentar as seguintes características: fácil manejo, prolificidade, docilidade, pequeno porte, baixo consumo alimentar, fisiologia conhecida e ciclo reprodutivo curto.

2. Objetivo

O presente estudo tem por finalidade avaliar os modelos lineares generalizados mistos, para caracterizar animais da espécie *Mus Musculus* produzidos pela FIOCRUZ.

3. Revisão de Literatura

A necessidade dos cálculos estatísticos se dá com a apresentação ou surgimento de um problema em que podemos aplicar esta ciência para analisar e interpretar seus resultados.

Segundo Sindelar et al. (2014), devemos sempre seguir etapas de forma que nos forneça todos os dados necessários, sempre de fontes seguras, para que nosso resultado seja o mais real possível. A primeira fase é a coleta de dados, ela pode ser “in loco”, por meio de censo, banco de dados e etc. Logo após, ocorre a classificação, que é identificar os dados que são necessários para a análise, por exemplo, idade, raça ou peso.

Na terceira etapa, apresentação e organização de dados, será a forma de organizar nossos dados, para facilitar a análise, pode ser por meio de tabelas, gráficos ou planilhas. Logo em seguida, já na fase final, iremos interpretar os resultados. A análise se dá através dos cálculos estatísticos e a interpretação será pelos resultados obtidos da amostra, que representa determinada população que estamos estudando (inferência estatística).

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 e cada dado ou informação tem seu modelo ideal de representação.

Segundo estudos realizados anteriormente, com o uso do modelo estatístico correto teremos uma interpretação mais verossímil dos dados e evitamos erros como utilização de dados com alta correlação, superestimação ou subestimação de dados.

3.1 Modelos lineares simples

Rencher (2000) conceituou esse modelo da seguinte forma: “nós nos preocupamos em modelar a relação entre duas variáveis, por exemplo, altura e peso, dose de uma droga e resposta, quantidade de adubo e produção de gramíneas”.

Para uma relação linear, usamos um modelo da forma:

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

No qual Y é a variável resposta, X é a variável explanatória e Ɛ é o termo de erro do modelo. 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.

Segundo Bussab (2017), faz-se necessário realizar algumas suposições referentes as variáveis aleatórias envolvidas. A primeira é que X é uma variável controlada, não sendo afetada por nenhum fator aleatório, sendo assim um fator fixo, sem erro. A segunda, os erros são não correlacionados, ou seja, para cada erro teremos um valor de X e, consequentemente, um valor de Y.

Em um modelo simples, temos somente uma variável X (independente) responsável por prever ou explicar a variável dependente (Y).

3.2 Modelos lineares generalizados mistos

A seleção dos modelos é de suma importância para toda pesquisa estatística, a escolha do modelo deverá atentar para o mais simples possível e que seus resultados descrevam bem e claramente os dados observados (CORDEIRO et al; 2013).

Nelder e Wedderburn (1972) mostraram, por meio de modelos, várias técnicas estatísticas que são desenvolvidas e normalmente estudadas de forma separadas e podem ser trabalhadas também de forma agrupada. A esse agrupamento denominaram de modelos lineares generalizados (MLG).

Cordeiro e Demétrio (2013) definiram os MLG como sendo formados a partir de três componentes, a variável resposta ou variável dependente, as variáveis explicativa ou variável independente e a função de ligação, cada um tem sua função, vejamos:

a) Variável resposta ou variável dependente ($Y_1, Y_2, ... Y_j$) = É definida assim que se especificam as medidas a serem utilizadas, podem ser contínuas ou discretas;

b) Variável explicativa ou variável independente ($X_1, X_2, ... X_i$) = 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.

A escolha de qual distribuição ou modelo estatístico utilizar irá depender da natureza dos dados, sejam eles discretos ou contínuos, seus intervalo de variação, conjunto dos reais, reais positivos e assimetria.

Sindelar (2014) definiu as variáveis aleatórias discretas e contínuas da seguinte forma: As variáveis aleatórias discretas 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.

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.

3.2.1 Distribuição de Poisson

Para essa distribuição, X é uma variável que 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 5 anos (SWUUNEY, 2013).

Os dados devem sempre atender as 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, onde seja t um valor positivo qualquer, T o tempo até a ocorrência do evento e X o número de ocorrências do evento no intervalo (0,t).

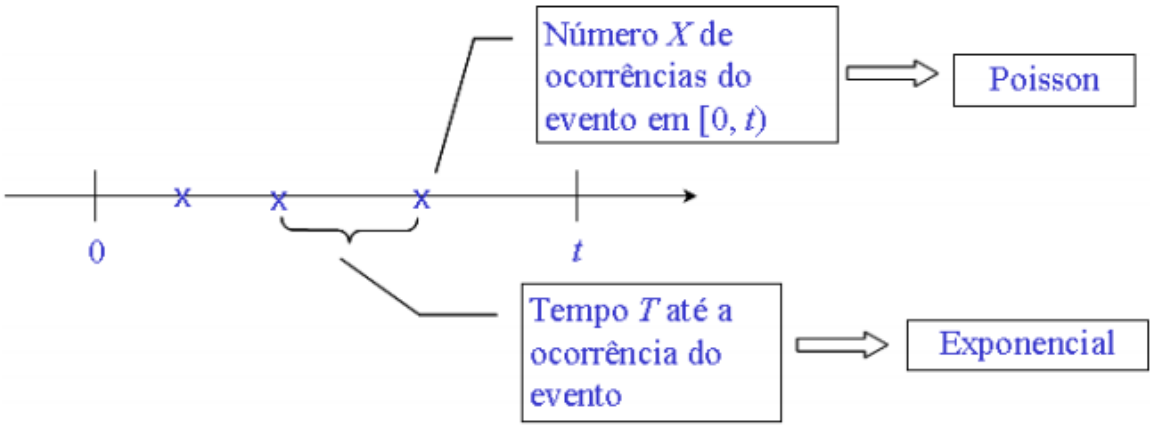

Figura 1: Distribuição de Poisson e distribuição exponencial.

Fonte: http://www.ufjf.br/clecio_ferreira/files/2013/10/Cap.-8-Principais-Variaveis-Aleatorias-continuas.pdf

3.2.2 Distribuições de probabilidade 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, 2013):

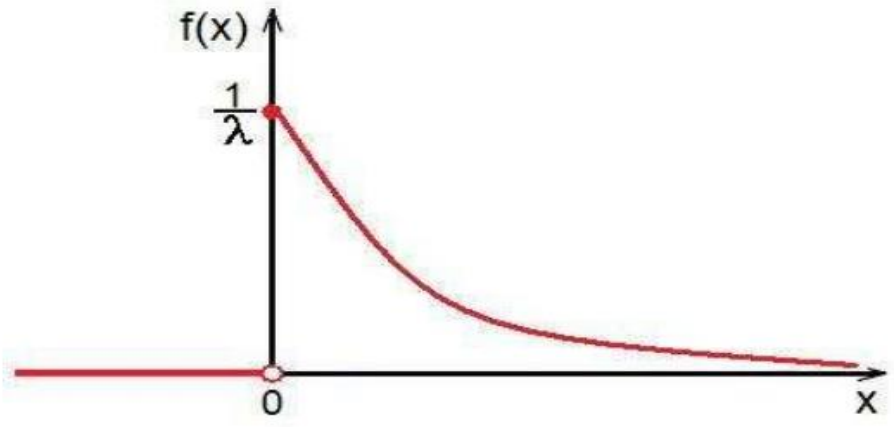

Figura 2: Gráfico de distribuição exponencial

Fonte: http://paginapessoal.utfpr.edu.br/ngsilva. Gráfico de distribuição exponencial.

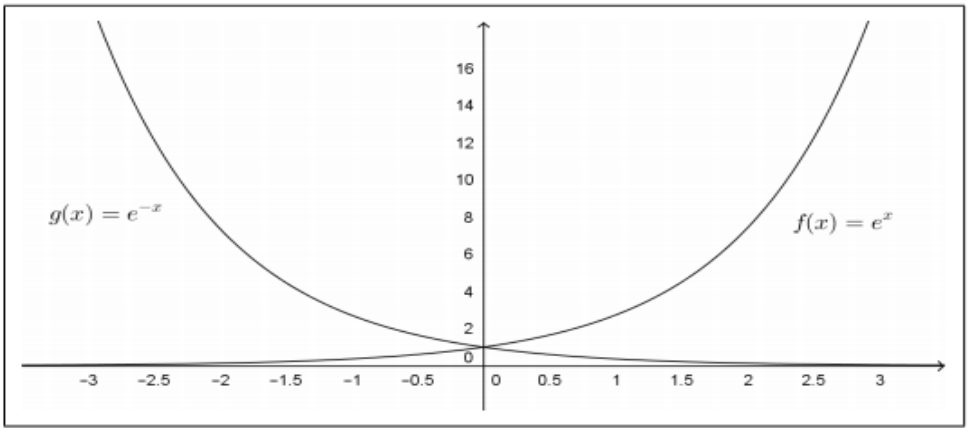

Figura 3: Gráfico de distribuição exponencial

Fonte: http://www.est.uff.br/index.php/ensino/material-didatico.

3.2.3 Distribuições de probabilidade 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 e etc. 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, ela surge quando indagamos o tempo necessário para obter um número específico de ocorrências do evento.

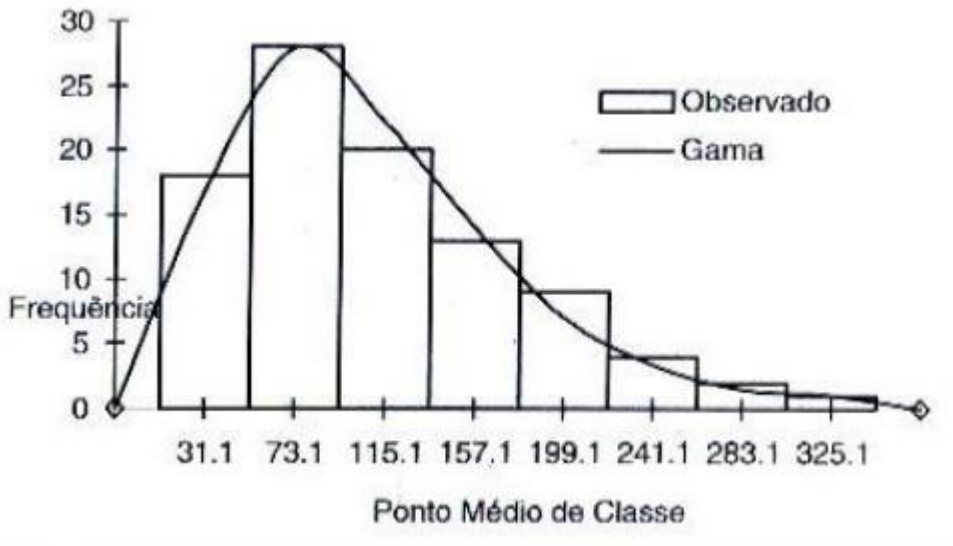

Figura 4: Gráfico de distribuição gama.

Fonte: www.bertolo.pro.br/FinEst/Estatistica/DistribuicaoContinua.pdf.

3.2.4 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, 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:

- 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. (STUDART, 2018).

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" (BERTOLO, 2018).

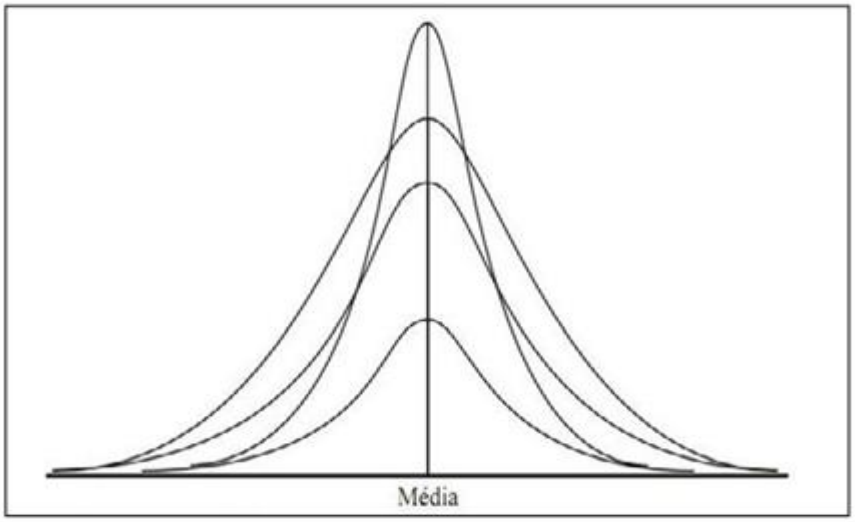

Figura 5: Gráfico de distribuição normal.

Fonte: Sindelar (2014). Gráfico de distribuição normal.

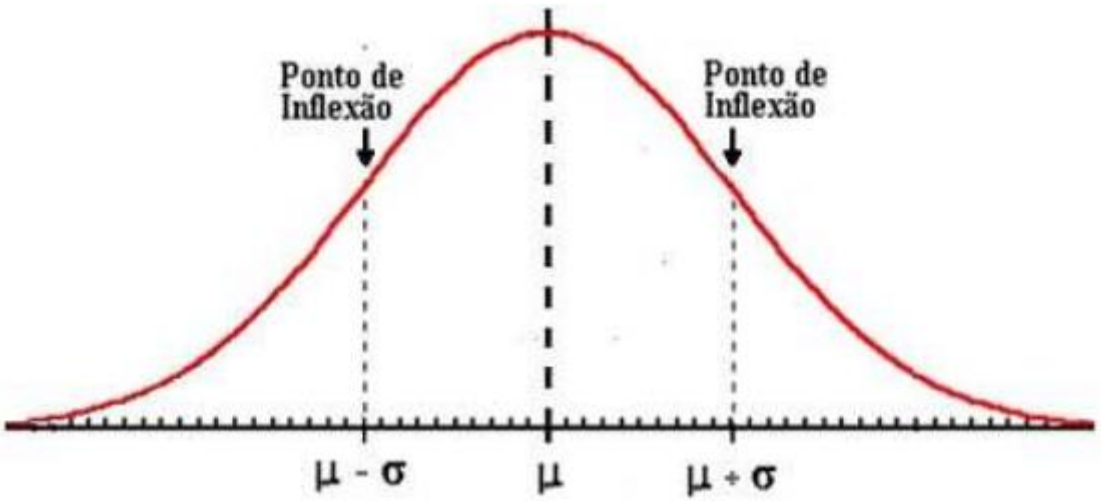

Figura 6: Gráfico de distribuição normal.

Fonte: http://paginapessoal.utfpr.edu.br/ngsilva. Gráfico de distribuição normal.

3.2.5 Distribuições de probabilidade 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. Qual a diferença entre a distribuição normal e log-normal? Ambas as formas de variabilidade são 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, 2010).

4. Ajuste do MLG

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

1. 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.

2. 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 (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.

3. 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 anda tem a ver com as observações em questão.

Os modelos lineares generalizados são bastante práticos, pois na etapa de formulação de modelos tem-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.

4.1 Qualidade de ajuste ou *goodness of fit (*GOF)

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,2017).

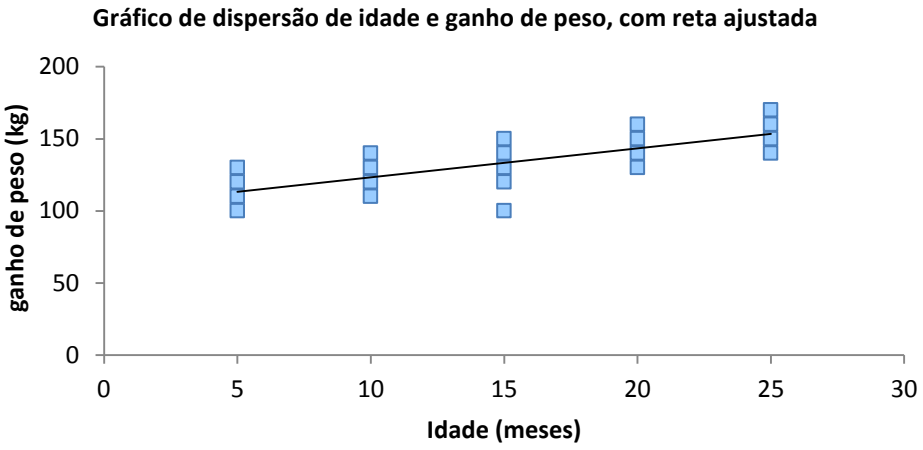

Figura **7**: Representação gráfica de ganho de peso e idade.

Fonte: Próprio autor.

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

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, Vonesh et al (1996) definiu 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_n) 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. Alguns autores preferem usar o R^2_a (coeficiente de determinação ajustado), a fim de evitar uma superestimação, definido como: (SWEENEY, 2013).

$$R^2_a = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

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

Assim como o coeficiente de determinação R^2, quanto maior o R^2_a, mais a variável resposta (Y) é explicada pela regressora X (portal action, 03.03.2018).

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).

5. Material e métodos

Serão utilizados 316 avaliações históricos dos animais da espécie *Mus Musculus,* mantidos no biotério da Universidade Estadual do Norte Fluminense (UENF) e seguindo as regras do comitê de ética da UENF. As características avaliadas serão: peso, número de animais nascidos, intervalos entre partos, entre outras. Com o auxílio de métodos estatísticos, modelos lineares generalizados mistos, que melhor as descrevem, associando a parâmetros que influenciam em tais características.

Os intervalos entre partos, número de animais nascidos e peso dos indivíduos das linhagens suíço, balb e nude serão avaliados com uso da teoria dos modelos lineares generalizados mistos pelo procedimento GLIMMIX do software SAS e a escolha da função distribuição estatística será pela macro %GOF do mesmo software.

6. Referências

SINDELAR, C. W; CONTO, S. M; AHLERT, L. **Teoria e prática em estatística para cursos de graduação.** 1 ed. Lajeado: Univates, 20A14.

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

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

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

http://www.bertolo.pro.br/FinEst/Estatistica/DistribuicaoContinua.pdf Acessado: 01.03.2018.

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.

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

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

VONESH, E. F.**Generalized linear and nonlinear models for correlated data.** 1 ed. Sas Institute Inc, Cary, USA : 2012.

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

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.

FERREIRA, C. S. **Cálculo de probabilidade I.** Prof UFJF, http://www.ufjf.br/clecio_ferreira/files/2013/10/Cap.-8-Principais-Variaveis-Aleatorias-continuas.pdf acessado 05.03.2018

ANDRADE, A; PINTO, S. C; OLIVEIRA, R. S. **Animais de laboratório: Criação e experimentação.** 20 ed. Rio de janeiro: 2002.

www.portalaction.com.br/analise-de-regressao/16-coeficiente-de-determinacao acessado 03.03.2018

7. Cronograma

Especificação da Atividade	Período: Abril/2017 a Março/2019			
	Sem. I	Sem. II	Sem. III	Sem. IV
Levantamento de dados da literatura	X	X		
Seleção, organização, tabulação e implementação de banco de dados		X	X	
Análise dos dados			X	X
Confecção Dissertação				X
Defesa de Dissertação				X