RODRIGO DE OLIVEIRA PACHECO

MODELOS DE EQUAÇÕES ESTRUTURAIS GENERALIZADOS MISTOS APLICADOS À AVALIAÇÃO GENÉTICA DE CODORNAS DE CORTE

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Zootecnia, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2014

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
P116m
2014

Pacheco, Rodrigo de Oliveira, 1986-
Modelos de equações estruturais generalizados mistos aplicandos à avaliação genética de codornas de corte / Rodrigo de Oliveira Pacheco. – Viçosa, MG, 2014.
x, 97f. : il. ; 29 cm.

Inclui apêndice.
Orientador: Robledo de Almeida Torres.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.89-93.

1. Condornas. 2. Genética animal. 3. Modelos Lineares (Estatística). 4. Modelo Recursivo. 5. Estrutura Causal. 6. Acurácia de Predição. I. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-graduação em Zootecnia. II. Título.

CDD 22. ed. 636.594

RODRIGO DE OLIVEIRA PACHECO

MODELOS DE EQUAÇÕES ESTRUTURAIS GENERALIZADOS MISTOS APLICADOS À AVALIAÇÃO GENÉTICA DE CODORNAS DE CORTE

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do programa de Pós-Graduação em Zootecnia, para a obtenção do título de *Doctor Scientiae*.

APROVADA: 25 de julho de 2014.

Antônio Policarpo Souza Carneiro

Moysés Nascimento

Fabyano Fonseca e Silva

Rafael Bastos Teixeira

Robledo de Almeida Torres
(Orientador)

Aos meus pais Catarina e Braz, pela força e incentivo;
ao meu irmão Luciano pela ajuda; à minha tia Catarina
por estar sempre ao meu lado; e aos meus sobrinhos
Miguel e Heitor. Dedico este trabalho.

"*O nosso destino está de acordo com os nossos méritos*" (Albert Einstein).

AGRADECIMENTOS

Acima de tudo agradeço à Deus, fonte de sabedoria. Aquele que me abençoou e deu forças para chegar onde eu nem sabia que poderia. Amigo fiel que nunca me desamparou.

Aos meus pais dou o maior obrigado que pode existir, eles representam o verdadeiro significado de família. Para minha mãe a maior incentivadora para chegar até aqui, sempre me lembro das palavras de coragem que me impulsionam, não tenho palavras para agradecer. Ao meu pai, que se tornou um ponto de apoio durante essa jornada, também agradeço imensamente, por a surpresa do seu suporte vale mais do que toda a riqueza deste mundo.

À Universidade Federal de Viçosa e ao Departamento de Zootecnia, agradeço a oportunidade.

Agradeço também à Coordenação de Aperfeiçoamento do Pessoal de Nível Superior, pela concessão da bolsa de doutorado e também pela bolsa do Programa de Doutorado Sanduíche no exterior.

Ao professor Robledo agradeço imensamente pela excelente orientação, não só para a vida acadêmica, mas também para a vida pessoal. Uma Pessoa com grande caráter e senso de humor. Com que se pode contar e lembrar sempre dos conselhos.

Ao professor Fabyano, sempre muito empolgado, a quem eu devo a ideia da realização deste trabalho e também a imensa ajuda para o doutorado sanduíche, uma das melhores experiências que vivi.

Agradeço a University of Wisconsin-Madison, ao professor Guilherme Rosa e ao Bruno Valente pelo suporte, atenção, disponibilidade orientar durante o período do doutorado sanduíche e pelo grande conhecimento agregado durante este período.

Aos membros da banca: Prof. Moysés Nascimento, Prof. Rafael Bastos e ao Prof. Antônio Policarpo Souza Carneiro, reforço o meu muito obrigado

Ao meu irmão Luciano, agradeço por sempre ser muito prestativo, sempre disposto a ajudar. À minha cunhada Rosamaria (Pepê) também deixo o meu muito obrigado. Não posso deixar de agradecer ao Miguel (meu bruxinho preferido) que mesmo dizendo que eu deveria morar mais perto, sem saber me deu muita força para seguir em frente.

Tia Catarina (tia mana) e tio Lécio, agradeço pelo carinho e apoio de pais, agradeço sinceramente pelos conselhos objetivos e também a todos os instantes que vibraram comigo a cada nova conquista.

À minha tia Iná por todos os elogios e considerações, por sua determinação, que serviu e servirá como exemplo, na qual me espelhei.

Aos amigos, companheiros de jornada, agradeço cada instante de convivência e que, de certo modo, carrego alguma coisa que aprendi com cada um: Mariele, Matilde, Carolina, Carla, Valentina, Rodrigo (Batata), Jeferson, Luciano, Joashllenny, Renata, Giovani e Aline. E também a todos os colegas de trabalho da Granja de Melhoramento de Aves. E também um agradecimento especial para Valentina Milani,

À minha namorada Daniela, seu extremamente grato pelo apoio, por ser uma grande companheira, por seu grande coração. Obrigado por estar ao meu lado.

E a todos que, direta ou indiretamente, colaboraram para a realização deste trabalho.

BIOGRAFIA

RODRIGO DE OLIVEIRA PACHECO, filho de Catarina de Fátima de Oliveira Pacheco e Braz Faraco Pacheco, nasceu em Petrópolis – RJ, em 6 de Fevereiro de 1986.

Iniciou o curso de graduação em Medicina Veterinária no Centro Universitário Serra dos Órgãos em fevereiro de 2003. De agosto a dezembro de 2004, foi monitor da disciplina de Histologia dos animais domésticos, na referida instituição.

Em julho de 2007 graduou-se em Medicina Veterinária pelo Centro Universitário serra dos Órgãos, em Teresópolis – RJ.

Em agosto de 2008 iniciou o mestrado em Zootecnia, sob a orientação do Professor Ricardo Frederico Euclydes, na Universidade Federal de Viçosa, em Viçosa – MG.

Iniciou o curso de Mestrado em Zootecnia em Agosto de 2008, sob a orientação do Professor Ricardo Frederico Euclydes, na Universidade Federal de Viçosa. Defendeu a dissertação de mestrado com o título de "Estudo genético da produção de ovos em uma linha de frango de corte por meio de análises multicaracterísticas e regressão aleatória" em julho de 2010.

Em agosto de 2010, iniciou o curso de Doutorado em Zootecnia, na área de melhoramento animal na Universidade Federal de Viçosa, sob a orientação do professor Robledo de Almeida Torres.

Em 25 de julho de 2014, submeteu-se aos exames finais de defesa de tese para a obtenção do título de *Doctor scientiae* em Zootecnia, na Universidade Federal de Viçosa.

RESUMO

PACHECO, Rodrigo de Oliveira, D.Sc., Universidade Federal de Viçosa, julho de 2014. **Modelos de Equações Estruturais Generalizados Mistos Aplicados à Avaliação Genética de Codornas de Corte**. Orientador: Robledo de Almeida Torres.

Os modelos de equações estruturais (MEE) são capazes de explicar a relação de causa e efeito entre as características. Os objetivos do presente estudo foram propor uma extensão destes modelos de equações estruturais sob o contexto dos modelos lineares generalizados mistos, a fim de consideram características com distribuição Poisson, e também avaliar o MEE generalizado sob o ponto de vista preditivo por meio da estimação da acurácia de predição. Foram avaliadas cinco características: Peso ao nascimento (PN), o peso aos 35 dias de idade (P35), a idade ao primeiro ovo (IPO), o peso médio dos ovos dos 42 aos 182 dias de idade (PMO) e o número de ovos produzidos dos 42 aos 182 dias de idade (NO). Foi determinada uma estrutura hipotética representando o relacionamento causal entre as características e os efeitos genéticos diretos sobre cada característica. Neste caso, foram desconsideradas as correlações genéticas entre estes efeitos genéticos. As equações do modelo estrutural recursivo foram estimadas seguindo formulação dos Modelos Lineares Generalizados Mistos, especificado. Os efeitos genéticos por si só não são suficientes para a determinação dos fenótipos, dessa forma esses efeitos devem ser associados com toda a informação que a estrutura causal pode oferecer em relação às características. Os valores da acurácia de predição foram considerados altos (0,87 – 0,99) para as características com exceção do número de ovos (0,355). De acordo com os valores de AIC e BIC, o modelo que considera o número de ovos (NO) com distribuição Poisson pode ser considerado o mais indicado do que o modelo que considera NO com distribuição normal. O modelo de equações estruturais lineares generalizados mistos é eficiente na descrição fenotípica e a avaliação do NO com distribuição Poisson foi melhor do que aquela que considera esta característica como sendo normalmente distribuída.

ABSTRACT

PACHECO, Rodrigo de Oliveira, D.Sc., Universidade Federal de Viçosa, July, 2014. **Generalized Mixed Structural Equation Models Applied to Meat-type Quails Genetic Evaluation**. Advisor: Robledo de Almeida Torres.

Structural Equation Models (SEM) are able to account the cause effect relationship among traits. The aims of this study were propose an extension for structural equation models in the generalized mixed linear models context, to consider traits with a Poisson distribution and evaluate the Generalized SEM under the predictive point of view by the accuracy of prediction estimation. Five meat-type quails traits were evaluated: Birth Weight (BW), Weight at 35 days of age (W35), Age at the first egg (AFE), Average Eggs Weight from 42 to 182 days of age (AEW), and number of eggs produced from 42 to 182 days of age (NE). I was determined a hypothetical structure representing the causal relationship among traits, and the direct genetic effects over each trait. In this case, the genetic correlations between these genetic effects were not considered. The equations of the recursive model were estimated following Generalized Linear Mixed Models formulation. The genetic effects per se were not sufficient for phenotype determination, for this reason, these effects should be associated with the whole information that the causal structure can offer in relation to traits. Accuracy of prediction values were considered high (0.87 – 0.99) for all traits but number of eggs (0.36). According to the AIC and BIC values, the model considering Number of eggs with Poisson distribution can be considered more indicated than the one considering number of eggs with a Gaussian distribution. Generalized Mixed Structural Equation model is efficient for phenotypic description and the evaluation considering NE with Poisson distribution was better than the one considering NE as a Gaussian distributed trait.

SUMÁRIO

INTRODUÇÃO

Modelos de Equações Estruturais (MEE) são modelos de regressão multivariados diferenciando-se da maioria dos outros modelos lineares multivariados pelo fato de que a variável resposta, em uma equação de regressão, pode aparecer como um preditor em uma nova equação. Neste tipo de modelo, as variáveis podem influenciar uma outra variável, de modo direto ou de modo intermediário através de outras varáveis (Fox, 2002). Desta forma, é possível estimar e testar as relações funcionais entre as características, que frequentemente não são reveladas por modelos lineares padrão e, o relacionamento causal entre as variáveis pode ser representado de forma significativa por estas equações estruturais (Fox, 2002; Rosa et al., 2011).

O pioneiro da modelagem de equações estruturais, por meio de análise de trilha, foi Sewall Wright, um dos pesquisadores mais influentes da história da genética quantitativa. Apesar disso, durante o século XX, estes modelos foram ignorados pelos pesquisadores na área da biologia. Em contra partida, ganharam importância em áreas como economia e ciências sociais. A adaptação dos MEEs para o contexto da genética foi feita por Gianola e Sorensen (2004) e a partir daí, foram aplicados e ampliados por vários autores que definiam a estrutura causal com base em crenças a respeito do sistema biológico estudado *a priori* (Valente, 2010; Valente et al., 2011).

A aplicação da metodologia de modelos de equações estruturais é justificada por Valente et al. (2013), que define duas vantagens de tais modelos: A primeira está relacionada com o fato de que as predições para diferentes cenários com modificações na rede de fenótipos não necessitam de dados obtidos de cenários extras, enquanto a abordagem através de modelos multicaracterísticos clássicos precisaria. E a segunda, é que, mesmo pequenas redes fenotípicas podem sofrer um enorme número de intervenções ou combinações de intervenções, de modo que a obtenção de dados para cada cenário possível e o ajuste de um MMC para cada cenário não é praticável.

É importante destacar que os, existem interpretações diferentes para os efeitos genéticos sob o contexto da modelagem de equações estruturais ou de modelos multicaracterísticos. Segundo Valente et al. (2013), nos MMCs os efeitos genéticos são expressos como um efeito genético global, equivalente a um conjunto de efeitos genéticos diretos e indiretos, afetando cada variável. Enquanto que nos MEEs, os efeitos genéticos são representados como agindo de forma direta sobre cada característica, sem a mediação de nenhuma característica contida no modelo, ou seja, sem efeitos indiretos. Uma

diferença fundamental entre os dois modelos é que um MEE não só descreve a distribuição dos dados, como também expressa o relacionamento causal entre as características.

Os modelos de equações estruturais são capazes de explicar a relação de causa e efeito entre as características. A metodologia proposta é uma extensão dos modelos de equações estruturais, sob o contexto dos modelos lineares generalizados mistos, permitindo que se considere na análise, características que tenham distribuição Poisson e ainda permite a avaliação individual de cada característica, mimetizando um modelo multicaracterístico através da inclusão do efeito dos pais das variáveis resposta em cada uma das equações.

O principal objetivo do presente estudo foi propor uma extensão dos Modelos de Equações Estruturais (MEE), sob o contexto dos modelos Lineares Generalizados Mistos, a fim de assumir características que não tenham distribuição Normal, tal como Poisson para número de ovos. Objetivou-se também avaliar o MEE generalizado sob o ponto de vista preditivo (estimação da acurácia de predição) por meio de análises de validação-cruzada.

REVISÃO DE LITERATURA

1. MODELOS LINEARES

Os modelos lineares ocupam uma posição de destaque em vários campos da ciência, incluindo o melhoramento animal. O objetivo desta metodologia é a investigação do relacionamento entre as variáveis estudadas, onde é possível considerar ou explicar a variação em uma variável em função de uma ou mais variáveis independentes (Searle, 1997; Haase, 2011).

1.1. Modelo Linear Univariado

No caso da regressão univariada simples, os modelos de regressão são aqueles que são limitados a uma única variável resposta (ou dependente) (Searle, 1997; Haase, 2011). Um modelo de regressão univariado simples, modelo este com uma única variável preditora (independente ou explanatória) pode ser expresso por:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + e\,. \tag{1}$$

Um modelo de regressão univariada múltipla, ou seja, um modelo mais complexo, com *q* variáveis preditoras é definido pela seguinte expressão:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_q \mathrm{X}_q + e\,. \tag{2}$$

Em (1) e (2), Y é um vetor coluna de variáveis respostas, explicado por meio da combinação linear dos coeficientes de regressão $\beta_0, \beta_1, \ldots, \beta_q$, das variáveis explanatórias $\mathrm{X}_0, \mathrm{X}_1, \ldots, \mathrm{X}_q$ e ainda por um termo de erro *e* (Littell et al., 2006; Haase, 2011).

Na notação matricial, (1) e (2) são dados por:

$$\mathbf{y}_{(n\times 1)} = \mathbf{X}_{(n\times q+1)}\boldsymbol{\beta}_{(q+1\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)}, \tag{3}$$

onde $\mathbf{y}_{(n\times 1)}$ é um vetor representando a variável resposta. X_j, $j = 1, 2, ..., q$, são as variáveis explanatórias alocadas em uma matriz de design $\mathbf{X}_{(n\times q+1)}$, de ordem $(n\times q+1)$ com n linhas, $i = 1, 2, ...,$ n, e $q+1$ colunas que captam as variáveis explanatórias. O "+1" na dimensão $q+1$ permite a uma unidade $X_0 \equiv 1$ (≡ significa "por definição igual a") estimar o intercepto do modelo. O vetor **β** da equação (3) é um vetor coluna dos coeficientes de regressão $(q+1\times 1)$ contendo uma linha pra cada uma das $q+1$ variáveis explanatórias. A expansão de (3) mostra os elementos contidos nas matrizes para um modelo univariado com $q+1$ variáveis explanatórias e é descrita como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_q \end{bmatrix}.$$

1.2. Modelo Linear Multivariado

Já os modelos lineares multivariados, são definidos por apresentar mais de uma variável resposta, que são incluídas simultaneamente na especificação do modelo. As variáveis explanatórias $X_0, X_1, ..., X_q$, podem ser iguais para o modelo univariado e para o modelo multivariado. Porém, o número de variáveis Y e *e* serão diferentes, com o número de colunas igual ao número dos coeficientes de regressão e ao número de termos de erro associados (Haase, 2011).

A representação do modelo de regressão múltipla multivariada é uma generalização de (3) dada por:

$$\mathbf{Y}_{(n\times p)} = \mathbf{X}_{(n\times q+1)}\mathbf{B}_{(q+1\times p)} + \mathbf{E}_{(n\times p)}. \qquad (4)$$

$\mathbf{Y}_{(n\times p)}$ é uma matriz na qual as linhas representam todas as *n* observações e as colunas contém $p > 1$ variáveis respostas, Y_k, para $k = 1, 2, ..., p$. Consequentemente, a ordem da matriz Y é $(n\times p)$. A estrutura da matriz de design $\mathbf{X}_{(n\times q+1)}$, não difere dos modelos univariados sendo idêntica a (3). A matriz $\mathbf{X}_{(q+1\times p)}$ da equação (4) é uma coleção aumentada dos coeficientes de regressão, uma linha para cada uma das $q+1$ variáveis

explanatórias e p colunas para acomodar as múltiplas variáveis respostas. E a matriz $\mathbf{E}_{(n\times p)}$ concentra os termos de erro, uma linha para cada um dos *n* indivíduos e uma coluna para cada uma das *p* variáveis resposta. Na expansão de (4) temos:

$$\begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix}$$

1.3. Estimação dos Parâmetros do modelo

Os modelos de equações (1) a (4) são funções de regressão com os parâmetros do modelo definidos no elementos de $\boldsymbol{\beta}_{(q+1\times 1)} = (\beta_0, \beta_1, \ldots, \beta_q)$ para os modelos uni variados e de $\mathbf{B}_{(q+1\times p)}$ para o caso multivariado. Para a q variável explanatória do modelo univariado da equação (3), o valor esperado da função para uma variável é dado por:

$$\mathrm{E}\left(\mathrm{Y} \mid \mathrm{X}_j\right) = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_q \mathrm{X}_q . \qquad (5)$$

Estes valores esperados são as médias da distribuição condicional de Y, chamada $\mu_{(\mathrm{Y}|\mathrm{X}_j)}$, para cada um dos valores de X_j. Os modelos lineares com duas variáveis explanatórias que a superfície de regressão definida por $\mathbf{X}\boldsymbol{\beta}$ seja um plano bidimensional com curvas parciais definindo os eixos de X em um gráfico. No caso de um modelo de regressão simples como em (1), o parâmetro β_0 define o valor esperado de $\mathrm{Y}|\mathrm{X} = 0$ e β_1 define a taxa de mudança esperada em Y por unidade modificada em X (Searle, 1997; Haase, 2011).

Dessa forma, todos os modelos podem ser descritos por

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} . \qquad (6)$$

A diferença entre Y e os valores esperados de Y são os erros de predição do modelo, representados por:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathrm{E}(\mathbf{y} \mid \mathbf{X}). \tag{7}$$

Os valores de β podem ser estimados pelo critério de mínimos quadrados, de modo que a discrepância entre as observações e os valores preditos pelo modelo sejam os menores possíveis. Quanto mais próximo os valores observados são dos valores ajustados pela regressão, melhor é o ajuste do modelo aos dados (Haase, 2011). Os valores de β são escolhidos para minimizar a soma de quadrados dos erros de predição, representado por:

$$\sum \boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \tag{8}$$

Substituindo as estimativas amostrais dos parâmetros da população $\hat{\boldsymbol{\beta}}_{(q+1\times 1)} = \left(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_q\right)$ em (8), pode ser mostrado que obtendo as derivações parciais de **ε**'**ε**, ajustando-as a zero e resolvendo o conjunto de equações simultâneas conduzem a uma solução dos coeficientes de regressão,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{Y}\right). \tag{9}$$

1.3.1. Métodos de estimação

a. <u>Mínimos quadrados ordinários</u>

Envolve a escolha de $\hat{\boldsymbol{\beta}}$ como o valor de **β** que minimiza a soma de quadrados dos desvios a partir dos seus valores esperados, ou seja, escolha de $\hat{\boldsymbol{\beta}}$ como **β** que minimize

$$\sum \left[y_i - E(y_i)\right]^2 = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right).$$

E o estimador resultante é como aquele apresentado anteriormente em (9) (Searle, 1997).

b. Mínimos Quadrados Generalizados

Ao assumir que a matriz de covariâncias de *e* é Var(*e*) = V, este método envolve a minimização de **V**

$$\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{'}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$$

em relação a **β** que leva a

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\right).$$

Os estimadores de mínimos quadrados ordinários e generalizados serão os mesmos quando a $\mathbf{V}=\sigma^2\mathbf{I}$ (Searle, 1997).

c. Máxima verossimilhança

A estimação pela máxima verossimilhança necessita de algumas suposições sobre a distribuição residual (que geralmente é normal) e a verossimilhança da amostra de observações representada pelos dados é então maximizada. Ao assumir que os erros têm distribuição normal com média zero e matriz de variância **V**, isto é, $e \sim N(\mathbf{0},\mathbf{V})$, a verossimilhança é

$$L=\left(2\pi\right)^{-\frac{1}{2}N}\left|\mathbf{V}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{'}\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\}.$$

Maximizando a equação em relação a de **β**, obtém-se o estimador da máxima verossimilhança de **β** que é

$$\tilde{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}\right),$$

tem a mesma forma do estimador de mínimos quadrados generalizados. Da mesma forma, quando $\mathbf{V} = \sigma^2\mathbf{I}$, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ (Searle, 1997).

1.3.2. Suposições necessárias para justificar a validade das estimativas de mínimos quadrados

Segundo Haase (2011) as suposições fornecem um grau de confiança na interpretação dos coeficientes bem como justificar a validade dos testes estatísticos. E dentre eles estão:

- O modelo é linear, a E(Y|X) está precisamente em uma linha reta;
- O modelo é especificado corretamente, nenhuma variável importante é omitida da análise;
- As X_j variáveis explanatórias são medidas sem erro.
- $E(\varepsilon) = 0$. Os erros do modelo de regressão são considerados como variáveis aleatórias com média zero.
- $\text{Cov}(\varepsilon_1,\varepsilon_j) = 0$, para $i \neq j$. Os erros são assumidos como independentes com covariância zero.
- $V(\varepsilon) = \sigma^2 I$, a variância dos erros e assumida como sendo constante. A variância é estimada pelo quadrado médio do resíduo, através da seguinte fórmula:

$$\hat{\sigma}^2 = \frac{\sum\left(Y - X\hat{B}\right)^2}{n - q_r - 1},$$

onde $n - q_r - 1$ representa os graus de liberdade do erro baseado em q_r variáveis explanatórias no modelo completo.

- $\varepsilon_i \sim N(X\beta,\sigma^2 I)$. Os erros do modelo são assumidos como normalmente distribuídos com média $X\beta$ e variância $\sigma^2 I$.

1.4. Definição das somas de quadrado das medidas de qualidade do ajuste

Em um modelo linear, a magnitude do relacionamento entre as variáveis resposta e as variáveis explanatórias é indicada pela Soma de quadrados residual (do resíduo ou dos erros)

$$SQ_{resíduo} = \sum \left(Y - X\hat{\beta} \right)^2 = \sum \varepsilon^2 = \hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}$$

e o coeficiente de correlação múltipla (R^2). Para alcançar cada uma dessas medidas é preciso que a variabilidade na variável resposta seja particionada em suas partes constituintes em relação a (3) (Haase, 2011). E a Soma de Quadrado pode ser dividia em

$$SQ_{Total} = SQ_{Modelo} + SQ_{resíduo}. \tag{10}$$

O vetor das estimativas de erro é dado por $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ e a soma de quadrado residual é definida por $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. Como a medida da qualidade do ajuste, $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ tem seus limites inferior e superior conhecidos, $0 \leq \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \leq SS_{Total}$, definindo a amplitude entre a não existência de relacionamento e um relacionamento perfeito. Considera-se $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ como uma medida ambígua como uma medida da força da associação entre as variáveis, a menos que a SQ_{Total} seja conhecido (Haase, 2011). A soma de quadrados total (média corrigida) é representada por

$$SQ_{Total} = \sum \left(Y - \bar{Y} \right)^2 = y'y - \bar{y}\bar{y}n,$$

onde $\bar{y}$ é um vetor das médias de Y repetidas $(n \times 1)$ n vezes. Redefinindo $y'y = (y'y - \bar{y}\bar{y}n)$ para ser a média corrigida SQ_{Total}, e redefinindo $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = (\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \bar{y}\bar{y}n)$ para representar a média corrigida do SQ_{Modelo}, A Soma de quadrados de (3) é dada por

$$\mathbf{y}'\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}. \tag{11}$$

É comum se basear no valor de R^2, que assume um valor no intervalo [0,1], como um índice de qualidade de ajuste. A SQ_{Total} é a variabilidade máxima disponível em Y, $SQ_{Resíduo}$ é a variabilidade em Y que não pode ser contabilizado pelo modelo, e a SQ_{Modelo} é a parte da variabilidade em Y que é considerada pelo modelo (Haase, 2011). A

proporção da variabilidade em Y que é contabilizada pelo modelo, R^2, é a medida escalonada da qualidade de ajuste e é computada como

$$R^2_{Y.\,X_1\,X_2\ldots X_q} = 1 - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\mathbf{y}'\mathbf{y}} \tag{12}$$

ou de forma mais comum

$$R^2_{Y.\,X_1\,X_2\ldots X_q} = 1 - \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}. \tag{13}$$

1.5. Teste de hipótese sobre os coeficientes de regressão e R^2

A confiabilidade de $\hat{\beta}$ ou R^2 dependem do conhecimento da variabilidade da amostragem da estatística e do teste estatístico para a avaliação das hipóteses sobre os parâmetros do modelo. Os dois métodos mais comuns incluem o teste F sobre os valores de R^2 e o teste *t* sobre o coeficiente de regressão do modelo onde $t = \sqrt{F}$. Um teste F genérico sobre gl_{mc} e gl_{mr}, graus de liberdade do modelo completo e do modelo restrito, respectivamente, é defino por

$$F_{(gl_{mc},gl_{mr})} = \frac{R^2_c - R^2_r}{1 - R^2_c} \cdot \frac{gl_{mr}}{gl_{mc}}. \tag{14}$$

No qual $gl_{mc} = q_{mc} - q_{mr}$ e $gl_{mr} = n - q_{mc} - 1$, sendo q_{mc} o número de covariáveis no modelo completo e q_{mr} o número de covariáveis no modelo restrito (Haase, 2011).

Para um teste com gl_{mc}, o teste *t* da hipótese $H_0 = \beta_j = k$ é expresso por:

$$t_{(gl_{mr})} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{QMR}{SQ_{X_j}}\left(\frac{1}{1 - R^2_{X_j}}\right)}}, \tag{15}$$

onde o QMR, o quadrado médio do resíduo é

$$\text{QMR} = \frac{\text{SQ}_{\text{resíduo}}}{\text{n} - \text{q}_\text{r} - 1},$$

e

$$\frac{1}{1 - \text{R}^2_{\text{X}_\text{j}}}$$

é o fator de inflação de variância (FIV), que ajusta para multicolinearidade entre os preditores lineares (Haase, 1997).

1.6. Teste de hipótese geral

Os testes de hipótese apresentados anteriormente são casos especiais do teste de hipótese linear geral, que se trata de um procedimento que cobre uma ordem de testes de hipóteses comuns e especializadas em casos modelos lineares uni e multivariados (Haase, 2011; Searle, 1997).

Quando há o interesse de se testar que todos os coeficientes de regressão do modelo completo são iguais a zero, com exceção do intercepto para o modelo (3), pode-se representar esta hipótese através da combinação linear dos parâmetros indicada por

$$\text{H}_0 : \mathbf{L}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{\text{qr}} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{\text{qr}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{16}$$

A matriz **L** é de ordem $\left(\text{c} \times \text{q}_\text{r} + 1\right)$ cuja função é identificar os coeficientes de interesse nas hipóteses. Outra hipótese poderiam envolver somente a estimação de um único parâmetro, por exemplo $\text{H}_0 : \beta_1 = 0$, ou algum subconjunto de parâmetros

$$H_0 : \begin{bmatrix} \beta_1 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Em geral qualquer hipótese pode ser definida como um produto de um vetor (ou matriz) de coeficientes de contrastes $\mathbf{L}_{(c \times q+1)}$, e do vetor de parâmetros $\boldsymbol{\beta}_{(q+1 \times 1)}$, da análise do modelo completo. Onde o subscrito c é o número de linhas em L, que é equivalente aos q_{mc}. Uma vez que a hipótese é especificada, as estimativas dos parâmetros podem ser substituídas em (21) para se obter a soma de quadrados das hipóteses

$$SQ_{\text{hipótese}} = \left(\mathbf{L}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{L}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{L}'\right)^{-1}\left(\mathbf{L}\hat{\boldsymbol{\beta}}\right) \tag{17}$$

e a $SQ_{\text{hipótese}}$ pode ser utilizada como o numerador do teste F (Haase, 2011, Searle, 1997).

$$F_{(gl_{mc}, gl_{mr})} = \frac{SQ_{\text{hipótese}}}{SQ_{\text{resíduo}}} \cdot \frac{gl_{mr}}{gl_{mc}} \tag{18}$$

Sob a suposição de que os erros seguem distribuição normal, F seguirá distribuição de $gl_{mc} = c$ e $gl_{mr} = n - q_{mc} - 1$ graus de liberdade.

A partir desta formulação, pode ser então promovida o teste de hipótese dos parâmetros do modelo como um todo, das contribuições individuais dos parâmetros ou da combinação destes (Haase, 2011).

2. MODELOS LINEARES MISTOS

Os Modelos Lineares Mistos são também conhecidos como Modelos Mistos ou modelos de efeitos mistos. São assim denominados, pois contém termos de efeito fixo e termos de efeito aleatório. (Costa, 2010).

Através da análise dos modelos mistos é possível estudar o comportamento individual e o comportamento médio, ao contrário dos modelos lineares clássicos (MLC) que ajustam o comportamento médio (Costa, 2010). Um motivo pelo qual um modelo misto pode ser adotado é necessidade de se realizar a predição de efeitos aleatórios, na presença de efeitos fixos (Martins et al., 1998).

2.1. Definição do Modelo

Em termos matriciais um modelo misto pode ser descrito na descrito como:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}, \tag{19}$$

onde **Y** é o vetor de observações, **X** é uma matriz de incidência dos efeitos fixos conhecida, **β** é o vetor de feitos fixos desconhecidos, **Z** é a matriz de incidência dos efeitos aleatórios conhecida, **u** é um vetor de efeitos aleatórios desconhecido e **ε** um vetor de erros.

As pressuposições das distribuições **y**, **u** e **ε** podem ser:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \left(\begin{bmatrix} \mathbf{X\beta} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} ZUZ' + R & ZU & R \\ UZ' & U & \phi \\ R & \phi & R \end{bmatrix} \right),$$

onde U é a matriz de covariâncias dos efeitos aleatórios presentes em **u** e R a matriz de covariâncias residuais (Martins et al., 1998).

Se cada uma das observações contiver mais de uma medida, a distribuição passa a ser multivariada e, se as medidas forem ordenadas dentro de cada observação de **y**, as matrizes U e R passam a ser:

- $U = A \otimes U_0$, sendo A a matriz de correlação entre os efeitos aleatórios, **u**, das n observações; e U_0 é a matriz de covariâncias entre os efeitos aleatórios na q medidas que compõem uma observação.
- $R = I_n \otimes R_0$, sendo I_n uma matriz identidade, e R_0 a matriz de covariâncias residuais entre as q medidas que compõem uma observação.

A derivação das equações de modelos mistos pode ser feita pela minimização do quadrado médio do erro ou pela maximização da função densidade de probabilidade conjunta de **y** e **u** (Martins et al., 1998). Resultando nas seguintes equações:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + U^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{o} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} X'R^{-1}\mathbf{y} \\ Z'R^{-1}\mathbf{y} \end{bmatrix}.$$

Segundo Martins et al. (1998), estas são as equações de modelos mistos que permitem obter as soluções para os efeitos fixos ($\boldsymbol{\beta}^o$) e as predições para os efeitos aleatórios ($\hat{\mathbf{u}}$). A solução do sistema pode ser obtida por absorção ou por obtenção da matriz inversa por partição. Nos dois casos, os resultados serão:

$$\beta^o = \left\{X'\left[R^{-1} - R^{-1}Z(Z'R^{-1}Z + U^{-1})^{-1}Z'R^{-1}\right]X\right\}^{-} X'\left[R^{-1} - R^{-1}Z(Z'R^{-1}Z + U^{-1})^{-1}Z'R^{-1}\right]y$$

$$\hat{u} = (Z'R^{-1}Z + U^{-1})^{-1}Z'R^{-1}(y - X\beta^o) \ .$$

2.2. Estimação dos parâmetros

Utilizando um modelo que ignora **u**, ou seja, (6)

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} ,$$

a solução $\boldsymbol{\beta}^o$, obtida pelas equações de modelos mistos, é também uma solução de mínimos quadrados generalizados (Martins et al., 1998). Sendo a variância de Y

$$\mathrm{Var}(Y) = V = ZUZ' + R ,$$

e a solução de mínimos quadrados generalizados para $\boldsymbol{\beta}$ é

$$\boldsymbol{\beta}^o = (X'V^{-1}X)^{-}X'V^{-1}y .$$

E pela segunda equação das equações de modelos mistos $\hat{\mathbf{u}}$

$$\hat{\mathbf{u}} = (Z'R^{-1}Z + U^{-1})^{-1}Z'R^{-1}(y - X\beta^o) .$$

Se

$$V^{-1} = R^{-1} - R^{-1}Z(Z'R^{-1}Z + U^{-1})^{-1}Z'R,$$

então, será verdade que $\boldsymbol{\beta}^o$, das equações de modelos mistos, é uma solução de mínimos quadrados generalizados para o modelo (6) que ignora **u**.

2.2.1. BLUE e BLUP

A variância de $\boldsymbol{\beta}^o$ é denotada por:

$$\mathrm{Var}(\boldsymbol{\beta}^o) = \left[X'R^{-1}X - X'R^{-1}Z(Z'R^{-1}Z + U^{-1})^{-1}Z'R^{-1}X \right]^{-} .$$

Para um dado conjunto de funções estimáveis, estabelecidas por uma matriz conhecida K, a variância de $K'\boldsymbol{\beta}^o$, o Melhor Estimador Não-Viesado (BLUE) de $K'\boldsymbol{\beta}$ é

$$\mathrm{Var}(K'\boldsymbol{\beta}^o) = K'\mathrm{Var}(\boldsymbol{\beta}^o)K .$$

De acordo com Martins et al. (1998), o Melhor Preditor Linear Não-Viesado (BLUP) pode ser, definido como resultado da regressão dos efeitos de um fator aleatório (u) em função das observações (y) corrigidas para os efeitos dos fatores fixos (Xβ), como dado na seguinte expressão:

$$\hat{\mathbf{u}} = UZ'(ZUZ' + R^{-1})(y - X\beta^o) \; = \; UZ'V^{-1}(y - X\beta^o) .$$

Observa-se que o termo $UZ'(ZUZ' + R)^{-1}$ é o conjunto de coeficientes de regressão em **u** em função de y, uma vez que DZ' é a matriz de covariâncias entre **u** e y. $(ZUZ' + R)^{-1}$ é a inversa da matriz de variância de y, enquanto o termo $(y - X\beta^o)$ contém os valores das observações y, corrigidas para os efeitos fixos Xβ (Martins et al., 1998).

Pelas equações de modelos mistos $\hat{\mathbf{u}}$ é

$$\hat{\mathbf{u}} = \left(Z'R^{-1}Z + U^{-1} \right)^{-1} Z'R^{-1} \left(y - X\beta^o \right) .$$

Então, se a igualdade

$$UZ'\left(ZUZ'+R\right)^{-1}\left(Z'R^{-1}Z+U^{-1}\right)^{-1}Z'R^{-1}$$

for verdadeira o **û**, obtido pelas equações de modelos mistos, é o BLUP de u.

A variância de **û** é dada por:

$$\mathrm{Var}(\hat{\mathbf{u}}) = UZ'\left[V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}\right]ZU'.$$

A correlação entre os valores reais e preditos é máxima.

2.3. Escolha do modelo

No caso de modelos que foram ajustados pelos métodos de Máxima Verossimilhança ou Máxima Verossimilhança Restrita existem diversos critérios que fornecem estatísticas que auxiliam na decisão da escolha de um modelo. Dois deles que são amplamente utilizados são o AIC (Akaike Information Criterion) (Akaike, 1973) e o BIC (Bayesian Information Criterion) (Schwarz, 1978) (Costa, 2010).

- AIC = $-2l + 2p$;
- BIC = $-2l + p.log(n)$.

Em que l é máximo da log-verossimilhança, p a quantidade de parâmetros e n o número de observações. Como critério de decisão, o melhor modelo é aquele que apresenta o menor valor em uma dessas duas estatísticas. O BIC é considerado mais consistente que o AIC, portanto, poderá ser ter maior peso na escolha do melhor modelo (Costa, 2010).

Outro critério que auxilia na escolha de modelos é o Teste da Razão de Verossimilhança (LRT), que se baseia na razão entres as verossimilhanças de dois modelos ajustados. Na prática, verifica se um modelo com menos parâmetros se ajusta tão bem quanto o modelo com a quantidade total de parâmetros. Segundo Pinheiro e Bates(2000)(2000)(2000)(2000)(2000)(2000)(2000)(2000)(2000)(2000), nos casos em que os parâmetros são estimados por Máxima Verossimilhança Restrita, o teste da Razão

de Verossimilhança só pode ser aplicado se os dois modelos foram ajustados pelo mesmo método e se os efeitos fixos têm a mesma estrutura.

A fórmula do teste da Razão de Verossimilhanças é dado por:

$$\mathrm{LRT} = 2\left[\log(\mathrm{L_{mc}}) - \log(\mathrm{L_{mr}})\right],$$

onde L_{mc} é o valor maximizado da log-verossimilhança sob o modelo completo e L_{mr} o valor maximizado da log-verossimilhança sob o modelo reduzido.

O teste da razão de verossimilhança apresenta distribuição assintótica de χ^2_r onde *r* é a diferença entre o número de parâmetros do modelo completo e do modelo restrito.

3. MODELOS LINEARES GENERALIZADOS

Os Modelos Lineares Generalizados (MLG) são extensões dos Modelo Lineares Clássicos (MLC) para casos em que os dados são independentes e as pressuposições do MLC são violadas. Estes modelos reúnem uma série de técnicas estatísticas que antes eram estudadas separadamente. Facilitando assim dessa forma que muitos problemas estatísticos de diferentes áreas pudessem ser formulados de uma forma unificada, como modelos de regressão (Demétrio, 2002; Littell et al., 2006).

3.1. Definição

3.1.1. Componentes de um MLG

Um vetor de observações **y** contendo *n* componentes é assumido como sendo uma variável aleatória **Y** cujos componentes são distribuídos de forma independe com média **μ**, e associado a ela um conjunto de variáveis explicativas $X_1, X_2, \ldots, X_p$. Para uma amostra de *n* observações $(y_i, \mathbf{x}_i)$ em que $\boldsymbol{x_i}=(x_1, x_2, \ldots, x_{pi})'$ é o vetor coluna de variáveis explicativas. O MLG envolve três componentes (McCullagh and Nelder, 1989; Demétrio, 2002):

a) **Componente sistemático**: as variáveis explicativas que entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \mathbf{x}_i'\boldsymbol{\beta} \text{ ou } \boldsymbol{\eta} = \mathbf{X}'\boldsymbol{\beta}$$

sendo $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)'$ a matriz do modelo, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p)'$ o vetor de parâmetros e $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n)'$ o preditor linear.

b) **Componente aleatório**: parte do modelo representada por um conjunto de variáveis aleatórias independentes $Y_1, Y_2, \ldots, Y_k$ provenientes de uma mesma distribuição da família exponencial na forma canônica com médias $\mu_1, \mu_2, \ldots, \mu_k$,

$$E(Y_i) = \mu_i, \quad i = 1,2,\ldots,n \quad \text{(McCullagh e Nelder, 1989).}$$

c) Funções de ligação

A função de ligação tem a função de relacionar o preditor linear η ao valor esperado de μ de um dado *y*, ou seja, é uma função que liga o componente aleatório ao componente sistemático

$$\eta_i = g(\mu_i),$$

sendo *g*(.) uma função monotônica, derivável (McCullagh e Nelder, 1989; Demétrio, 2002).

Devido a função de ligação ser monotônica, o relacionamento entre μ e η (bem como entre μ e β) pode ser expresso em termos da função de ligação ou o inverso da função (Littell et al., 2006):

$$g(\boldsymbol{\mu}) = \mathbf{X}'\boldsymbol{\beta} = \boldsymbol{\eta}$$

$$\boldsymbol{\mu} = g^{-1}(\mathbf{X}'\boldsymbol{\beta})$$

A primeira forma é ilustrativa pois enfatiza que os MLGs utilizam transformações da média; estes não envolvem transformações dos dados. A segunda é útil pois mostra como as predições da média podem ser obtidas, seguindo a estimação de **β**: calcula a estimativa de **η** e aplica a função de ligação inversa,

$$\hat{\boldsymbol{\mu}} = g^{-1}(\hat{\boldsymbol{\eta}}) = g^{-1}\left(\mathbf{X}'\hat{\boldsymbol{\beta}}\right).$$

Se a função de ligação é escolhida de tal forma que $g(\mu_i) = \theta$, o preditor linear modela diretamente o parâmetro canônico e tal função de ligação é chamada ligação canônica. Isto resulta, frequentemente, em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão, além de vantagens teóricas em termos da existência de um conjunto de estatísticas suficientes para os parâmetros β's e alguma simplificação no algoritmo de estimação (Demétrio, 2002).

Para o MLC, a função de ligação é chamada identidade, pois o preditor linear é igual à média. Essa função de ligação é adequada no sentido em que ambos, η e μ, podem assumir valores na linha real. Entretanto, quando se trabalha com dados de contagem e a distribuição em questão é a Poisson, $\mu > 0$, dessa forma a função identidade não é indicada, porque, por exemplo, η pode ter um valor negativo, dependendo dos valores obtidos para $\hat{\boldsymbol{\beta}}$, enquanto μ não. Modelos para contagem baseados na independência dos dados conduzem naturalmente a um efeito multiplicativo. Nesse caso, a função de ligação pode ser expressa ligação log, onde $\eta = \ln \mu$, com inversa $\mu = e^{\eta}$. Logo, a contribuição dos efeitos aditivos a η se tornam a contribuição dos efeitos aditivos a μ e, μ é necessariamente positivo (McCullagh e Nelder, 1989; Demétrio, 2002).

São três as funções de ligação consideradas principais. Elas são conhecidas por:

1- *Logit*

$$\eta = \log\{\mu / 1 - \mu\};$$

2- Probit

$$\eta = \Phi^{-1}(\mu),$$

onde Φ(.) é a função da distribuição Normal cumulativa; e

3- Complementar log-log

$$\eta = \log\left\{-\log\left(1-\mu\right)\right\}.$$

A família das ligações é sempre importante, pelo menos para observações com média positiva. Esta família pode ser especificada ou por

$$\eta = \left(\mu^{\lambda} - 1\right)/\lambda$$

com valor limite

$$\eta = \log\mu\,;\ \text{como}\ \ \lambda \to 0\,,$$

ou por

$$\eta = \begin{cases} \mu^{\lambda}; & \lambda \neq 0, \\ \log\mu; & \lambda = 0. \end{cases}$$

A primeira forma tem a vantagem de uma transição mais suave conforme λ passa por zero, mas quando $\lambda = 0$, deve-se optar por uma forma especial.

3.1.2. Predições de medias pela função de ligação inversa

A **função de ligação inversa** é definida como $g^{-1}(\eta) = \mu$. A ligação inversa é utilizada para obter os valores preditos de μ a partir do vetor de estimativas β (Littell et al., 2006).

$$\hat{\mu} = g^{-1}\left(\mathbf{x}'\hat{\boldsymbol{\beta}}\right)$$

Para a distribuição normal, $g^{-1}(\mathbf{x'\beta}) = (\mathbf{x'\beta})$, desde que $\mu = \eta$. Para a distribuição Poisson com ligação canônica, $\eta = \log(\lambda)$ consequentemente $\lambda = g^{-1}(\mathbf{x'\beta}) = \exp(\mathbf{x'\beta})$. Para a distribuição binomial, $\eta = \log[\pi/(1-\pi)]$ e consequentemente $\pi = g^{-1}(\mathbf{x'\beta}) = \exp(\mathbf{x'\beta})/ [1 + \exp(\mathbf{x'\beta})]$ (Littell et al., 2006).

3.1.3. Estatísticas suficientes e ligações canônicas

Cada uma das distribuições da família exponencial tem uma função de ligação especial para a qual existe uma estatística suficiente de dimensão igual a β no preditor linear. Estas funções de ligação são chamadas de *ligações canônicas*, e ocorrem quando

$$\theta = \eta,$$

onde θ é o parâmetro canônico (McCullagh e Nelder, 1989). As funções canônicas para distribuições da família exponencial estão descritas na Tabela 1:

Tabela 1: Funções de ligação canônicas

Distribuição	Ligação canônica	
Normal	Identidade:	$\eta = \mu$
Poisson	Logarítmica:	$\eta = \log \mu$
Binomial	Logística:	$\eta = \log\{\pi/(1-\pi)\} = \log\{\mu/(m-\mu)\}$
Gamma	Recíproca:	$\eta = \mu^{-1}$
Normal inversa	Recípocra[2]:	$\eta = \mu^{-2}$.

Adaptado de (McCullagh e Nelder, 1989).

3.1.4. Distribuições de Probabilidade

Os elementos essenciais para estimar **β** no MLG são:

- A **função de ligação**, que determina **η** e **D**.
- A **distribuição de probabilidade**, que determina a média **μ** e a variância Var[**Y**] = **R**.

O processo de escolha de uma função de ligação e as estruturas da média e da variância pode ser melhor compreendido examinando a distribuição de probabilidade,

mais especificamente, a função de verossimilhança. Considere as três distribuições da família exponencial que são as mais amplamente utilizadas, que são a binomial, Poisson e normal. Estas distribuições apresentam as características essenciais do MLG (Littell et al., 2006).

Binomial

Para a distribuição binomial com *n* tentativas e uma probabilidade de sucesso *p*, a forma da função de distribuição de probabilidade é dada por

$$f\left(y\right)=\binom{n}{y}p^{y}\left(1-p\right)^{n-y}.$$

A função de verossimilhança para a distribuição binomial é:

$$L\left(p\,|\,n,y\right)=\binom{n}{y}p^{y}\left(1-p\right)^{n-y}.$$

Esta função envolve o parâmetro *p*, dado os dados (*n* e *y*). Os dados discretos e a estatística *y* são conhecidos.

A função log-verossimilhança para o modelo binomial é dessa forma:

$$l(L)=\log\binom{n}{y}+y\log(p)+(n-y)\log(1-p)$$

Poisson

A probabilidade para a distribuição Poisson é

$$f\left(y\right)=\frac{\lambda^{y}e^{-\lambda}}{y!}.$$

Para a distribuição Poisson a função de verossimilhança é:

$$L(y) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$
$$= \frac{\lambda \sum_{i=1}^{n} y_i e^{-n\lambda}}{\prod_{i=1}^{n} y_i!},$$

e a função log-verossimilhança é

$$l(L) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \right)$$
$$= \sum_{i=1}^{n} \log\left(\frac{\lambda \sum_{i=1}^{n} y_i e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} \right)$$
$$= \sum_{i=1}^{n} y_i \log(\lambda) - \sum_{i=1}^{n} \log(y_i!) - n\lambda .$$

A média e variância para a distribuição Poisson são ambas iguais a λ.

Tradicionalmente, contagens seguem uma distribuição Poisson. Contudo, a distribuição Poisson assume que a média e a variância são iguais. Pesquisas recentes sugerem que os dados de contagem são tipicamente superdispersos, ou seja, a variância é maior (às vezes consideravelmente maior) do que a média. Como alternativa a este problema, a distribuição binomial negativa, frequentemente fornece melhores modelos de variação dos erros em um modelo linear (McCullagh e Nelder, 1989).

Normal

Para a distribuição normal, a função densidade de probabilidade é:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right],$$

e a função de verossimilhança é:

$$L\left(\mu,\sigma^2;y\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(y_i-\mu\right)^2}{2\sigma^2}\right]$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left[-\frac{\sum_i\left(y_i-\mu\right)^2}{2\sigma^2}\right].$$

Dessa forma a função log-verossimilhança para a distribuição normal é:

$$l(L)=-\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y-\mu\right)^2$$

onde μ é média e σ^2 a variância.

Características em comum

As funções log-verossimilhança para as três distribuições supracitadas têm a seguinte forma em comum:

$$l\left(\theta,\phi;y\right)=\frac{y\theta-b\left(\theta\right)}{a\left(\phi\right)}+c\left(y,\phi\right).$$

onde:

θ é o parâmetro natural; e

ϕ é o parâmetro de escala.

θ é uma função da média. Esta função é denotada por $\theta(\mu)$. Da mesma maneira a variância pode ser expressa como uma função da média e $a(\phi)$. Especificamente,

$$\mathrm{Var}[\mathrm{Y}] = \mathrm{V}(\mu)\, a(\phi),$$

onde V(μ) é a **função de variância**. Se $a(\phi) = 1$, então a função de variância também é a variância de uma observação, que é o caso para as distribuições de binomial e Poisson. Para tais distribuições, as funções podem são apresentadas na Tabela 2.

Tabela 2 – Funções para as distribuições Binomial, Poisson e Normal

	Binomial/*n*	**Poisson**	**Normal**
Média	π	λ	μ
$\theta(\mu)$	$\log[\pi/(1-\pi)]$	$\log(\lambda)$	μ
$a(\phi)$	$1/n$	1	σ^2
$V(\mu)$	$\pi(1-\pi)$	λ	1
Var[Y]	$\pi(1-\pi)/n$	λ	σ^2

(Littell et al., 2006).

Deve-se notar que para as distribuições Poisson e binomial, a média também aparece na função de variância, que é uma generalidade das distribuições da família exponencial. Na distribuição normal, a independência da média e da variância é uma esperança, isto é, é possível determinar o grau de variabilidade independentemente da média, somente para a distribuição normal. No caso das outras distribuições da família exponencial, o conhecimento da média também determina o grau de variabilidade, pelo menos até a função $a(\phi)$. Para a formulação e estimação, isto tem consequências de longo alcance (McCullagh and Nelder, 1989; Littell et al., 2006).

3.2. Ajuste do Modelo

McCullagh e Nelder (1989) destacam três processos no ajuste dos Modelos Lineares Generalizados: a seleção do modelo, a estimação cos parâmetros e a predição dos valores futuros.

3.2.1. Seleção de modelo

Os modelos selecionados para ajustar os dados são escolhidos dentro de uma determinada classe e, esta classe tem que ser relevante ao tipo de dados que estão sendo estudados para que esse processo de ajuste seja útil. Assim como na regressão clássica, uma característica importante dos MLGs é que estes assumem independência das observações (ou que não sejam correlacionadas) (McCullagh e Nelder, 1989).

A escolha da escala dos dados é um importante aspecto da seleção do modelo. Uma escolha comum ao se analisar *Y* seria utilizar a escala normal ou log *Y*. A escala mais adequada vai depender do propósito pelo qual uma escala está sendo utilizada (McCullagh e Nelder, 1989).

Na análise de regressão clássica uma boa escala deve combinar variância constante, normalidade aproximada dos erros a aditividade dos erros sistemáticos. Com a introdução dos MLGs, os problemas de escala são grandemente reduzidos. A normalidade e variância constante não são mais exigidos, embora que, a maneira como a variância depende da média deve ser conhecida. Nos MLGs, a aditividade é, devidamente, postulada como uma propriedade das respostas esperadas (McCullagh e Nelder, 1989).

Na seleção do modelo existe um problema que é a escolha das *x*-variáveis (ou covariáveis) a serem incluídas na parte sistemática do modelo. Agrupado em torno do 'melhor' modelo existirão um grupo de alternativas quase tão boas que não são estatisticamente distinguíveis (McCullagh e Nelder, 1989).

3.2.2. Estimação dos parâmetros

Uma vez feita a seleção de um determinado modelo, é preciso estimar os parâmetros e avaliar a precisão das estimativas. No caso dos MLGs, a estimação segue por definição uma medida de qualidade do ajuste entre os dados observados e os valores ajustados pelo modelo. As estimativas dos parâmetros são os valores que minimizam o critério de qualidade de ajuste (McCullagh e Nelder, 1989).

De acordo com McCullagh e Nelder (1989), Se $f(y; \theta)$ for a função densidade ou distribuição de probabilidade para a observação y dados os parâmetros θ, então o log da

verossimilhança, expressada como uma função do parâmetro da média, $\mu = E(Y)$, é apenas

$$\ell(\mu; y) = \log f(y; \theta).$$

O log da verossimilhança baseado em um conjunto de observações independentes $y_1, \ldots, y_n$, é apenas a soma das contribuições individuais, de maneira que

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_i \log f_i(y_i; \theta_i),$$

onde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Note que a função densidade $f(y; \theta)$ é considerada como uma função de y para θ fixo, enquanto que o log da verossimilhança é considerado essencialmente como uma função de θ para o dado particular y observado (McCullagh e Nelder, 1989).

3.2.3. Predição de valores

McCullagh e Nelder (1989), afirmam que neste tópico se encontra a preocupação com as respostas para questões do tipo ‘e se’, que podem ser colocadas de acordo com a análise estatística. Par ser útil, o valor predito precisa ser acompanhado por medidas de precisão, que são calculadas de forma ordinária sobre a suposição que a construção que produziram os dados se mantém constantes, e que o modelo utilizado na análise é substancialmente correto.

3.3. Estimação do vetor de parâmetros β

De acordo com Nelder e Wedderburn (1972), estimativas de máxima verossimilhança para **β** podem ser obtidas de forma iterativa resolvendo

$$\mathbf{X'WX\beta} = \mathbf{X'Wy}^*$$

onde,

$\mathbf{W} = \mathbf{D}'\mathbf{R}^{-1}\mathbf{D}$;

$\mathbf{y}^{*}\boldsymbol{\eta} = \hat{\mathbf{y}} + \left(\boldsymbol{\mu}\mathbf{D}'\right)^{-1}$;

$\mathbf{D}\boldsymbol{\mu} = \partial\boldsymbol{\eta} / \partial$;

$\mathbf{R} = \mathrm{Var}[\mathbf{Y}]$; e

$\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}]$.

$\mathbf{y}^{*}$ é a variável resposta que está sendo estudada.

Na prática, as estimativas de **D** e **R** são utilizadas no lugar de **D** e **R**. para os modelos lineares clássicos(MLCs), $\boldsymbol{\eta} = \mathrm{E}[\mathbf{Y}] = \boldsymbol{\mu}$ e por isso **D** = **I**. Consequentemente, a solução para **β** reduz os mínimos quadrados generalizados. Isto é, $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y}$.

3.4. Qualidade do ajuste

O objetivo é determinar quantos termos são necessários na estrutura linear para uma descrição razoável dos dados. Quando se tem um pequeno número de covariáveis (varáveis explanatórias), isso pode levar a um modelo de fácil interpretação, porém que se ajuste de forma pobre aos dados. Já quando se tem um grande número de covariáveis, isso pode gerar um modelo que explique bem os dados, mas há um aumento na complexidade da interpretação. Então, o que se é um modelo intermediário (Demétrio, 2002).

Dadas *n* observações, a elas podem ser ajustados modelos contendo até *n* parâmetros. O modelo mais simples é o **modelo nulo** tem apenas um parâmetro, que apresenta um μ comum a todos os *y*'s. Assim, o modelo nulo deposita toda a variação entre os *y*'s no componente aleatório. A matriz do modelo, então, reduz-se a um vetor coluna, formado de 1's. em situação oposta, encontra-se está o **modelo completo** ou **saturado** que possui *n* parâmetros, sendo um para cada observação, e os μ's derivados dele se igualam exatamente aos dados. Este por sua vez, atribui toda a variação nos *y*'s ao componente sistemático (McCullagh e Nelder, 1989; Demétrio, 2002).

Na prática o modelo nulo, geralmente, é demasiado simples e o modelo completo é não informativo, pois não resume os dados, mas sim os repete como um todo. Contudo, o modelo completo dá uma base para as medidas de discrepância para um modelo alternativo com *p* parâmetros (McCullagh, 1989).

Existem ainda dois outros modelos limitantes, porém, menos extremos. É necessário que certos parâmetros estejam no modelo, como por exemplo, os totais marginais fixados em tabelas de contingência. O **modelo maximal** é o modelo que contém o maior número de termos que podem ser considerados. Por outro lado, o **modelo minimal** é aquele que contém o menor número de termos necessários para o ajuste. Os termos desses modelos extremos são obtidos, geralmente, através de interpretações da estrutura dos dados, feitas *a priori* (Demétrio, 2002).

Em geral, trabalha-se com modelos encaixados e o conjunto de matrizes dos modelos pode, então, ser formado pela adição sucessiva de termos ao modelo minimal até se chegar ao modelo maximal. Qualquer modelo com *p* parâmetros linearmente independentes, situado entre os modelos minimal e maximal**,** é chamado **modelo corrente** ou **modelo sob pesquisa**. O problema é determinar a utilidade de um parâmetro extra no modelo corrente (sob pesquisa) ou, então, verificar a falta de ajuste induzida pela omissão dele. A fim de discriminar entre modelos, medidas de discrepância devem ser introduzidas para medir o ajuste de um modelo (Demétrio, 2002).

3.4.1. Deviance

Nelder & Wedderburn (1972) propuseram como medida de discrepância entre os modelos, a *deviance* (traduzida como **desvio**), em que o log da verossimilhança em termos de valor médio do parâmetro $\boldsymbol{\mu}$ ao invés do parâmetro canônico $\boldsymbol{\theta}$. Seja $\ell(\hat{\mu}, \phi; y)$ o log da verossimilhança maximizada sobre $\boldsymbol{\beta}$ para um valor fixo do parâmetro de dispersão ϕ. A máxima verossimilhança alcançável em um modelo completo com *n* parâmetros é $\ell(\mathbf{y}, \phi; \mathbf{y})$.

A *deviance* é proporcional a duas vezes a diferença entre o logaritmo da verossimilhança do modelo completo e do logaritmo da verossimilhança do modelo que está sendo estudado (McCullagh e Nelder, 1989). Com expressão dada por:

$$D(\hat{\mu}; y) = 2\left[l(y; y) - l(\hat{\mu}; y) \right],$$

onde,

$l(y; y)$ é o valor do log da verossimilhança calculado em $\mu = y$ (modelo saturado).

$l(\hat{\mu}; y)$ é o valor do log da verossimilhança para o modelo corrente.

A *deviance* é uma generalização da soma de quadrados residuais (SQR) na análise de variância e da razão de verossimilhança χ^2 em tabelas de contingência. A *deviance* é igual à SQR para modelos normais e a razão de verossimilhança χ^2 é igual a *deviance* para modelos Poisson. A *deviance* pode ser utilizada na avaliação da qualidade do modelo e nos testes de hipótese (Littell et al., 2006).

As formas da *deviance* para as distribuições da família exponencial são apresentadas por McCullagh e Nelder (1989). Com somatório com índice $i = 1, \ldots, n$:

Normal	$\sum (y - \hat{\mu})^2$
Poisson	$2\sum \{ y - \log(y / \hat{\mu}) - (y - \hat{\mu}) \}$
Binomial	$2\sum \{ y - \log(y / \hat{\mu}) + (m - n) \log[(m - y) / (m - \hat{\mu})] \}$
Gamma	$2\sum \{ -\log(y / \hat{\mu}) + (y - \hat{\mu}) / \hat{\mu} \}$
Normal inversa	$\sum (y - \hat{\mu})^2 / \hat{\mu}^2 y$

Quando ϕ não é conhecido, este pode ser estimado e utilizado para calcular a **scaled deviance**, que é definida por

$$D^*(\hat{\boldsymbol{\mu}}; \mathbf{y}) = D(\hat{\boldsymbol{\mu}}; \mathbf{y}) / \hat{\phi},$$

em que D^* é chamada de *scaled deviance* (McCullagh e Nelder, 1989). As funções *scaled deviance* para as distribuições da família exponencial estão descritas na Tabela 3.

Tabela 3: Funções *scaled deviance* para algumas distribuições

Distribuição	*Scaled deviance*
Normal	$S_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} w_i (y_i - \hat{\mu}_i)^2$
Binomial	$S_p = 2\sum_{i=1}^{n} w_i \left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\ell n\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right]$
Poisson	$S_p = 2\sum_{i=1}^{n} w_i \left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$
Binomial negativa	$S_p = 2\sum_{i=1}^{n} w_i \left[y_i \ell n\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i + k)\ell n\left(\frac{y_i + k}{\hat{\mu}_i + k}\right)\right]$
Gama	$S_p = 2\nu\sum_{i=1}^{n} w_i \left[-\ell n\left(\frac{y_i}{\hat{\mu}_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right]$
Normal Inversa	$S_p = \frac{1}{\sigma^2}\sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$

(Demétrio, 2002).

A *deviance* é sempre maior do que ou igual a zero, e à medida que são adicionadas covariáveis no componente sistemático, a deviance decresce até se tornar zero (modelo saturado). Quanto melhor for o ajuste do modelo aos dados tanto menor será o valor de D^*. Assim, um modelo bem ajustado aos dados com uma verossimilhança grande tem uma *deviance* pequena. Uma maneira de se conseguir a diminuição da *deviance* é aumentar o número de parâmetros, o que, porém, significa um aumento do grau de complexidade na interpretação do modelo. Na prática, procuram-se modelos simples com *deviance* moderada, situados entre os modelos mais complicados e os que se ajustam mal aos dados (Demétrio, 2002).

Para testar a adequação de um modelo linear generalizado, tem-se

$$D^* = \frac{D}{\sigma^2} \sim \chi^2_{n-p} \ ,$$

com $(n - p)$ graus de liberdade, e assumindo σ^2 conhecida (Demétrio, 2002).

Assumindo-se que o modelo é verdadeiro, para a distribuição binomial, quando n é fixo e $m_1 \to \infty$, $\forall i$ e para a distribuição de Poisson, quando $\mu_i \to \infty$, $\forall i$, tem-se que:

$$D^* = D \sim \chi^2_{n-p} \ .$$

Na prática, contenta-se em testar um modelo linear generalizado, sem muito rigor, comparando-se o valor D^* com os percentis da distribuição χ^2_{n-p}. Assim, nos casos em que é possível a aproximação de uma χ^2_{n-p}, tem-se que se

$$D^* \leq \chi^2_{n-p;\alpha}.$$

Pode-se considerar que existem evidências, a um nível aproximado de α = 100% de probabilidade, que o modelo proposto está bem ajustado aos dados (Demétrio, 2002).

3.4.2. Estatística $\chi 2$ generalizada de Pearson

Outra medida importante da discrepância do ajuste de um modelo é a estatística χ^2 generalizada de Pearson, denotada por

$$\chi^2 = \sum_{i=1}^{n} w_i \frac{\left(y_i - \hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)},$$

onde $V(\hat{\mu}_i)$ é a função de variância estimada sob o modelo que está sendo ajustado aos dados. Para a distribuição normal, o χ^2 é igual à soma de quadrados do resíduo e

$$\frac{\chi^2}{\sigma^2} \sim \chi^2_{n-p}$$

(McCullagh e Nelder, 1989; Demétrio, 2002).

Para as distribuições binomial e de Poisson, em que ϕ = 1, é a estatística original de Pearson, escrita na forma

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i},$$

sendo O_i a frequência observada e E_i a frequência esperada (Demétrio, 2002).

Para as distribuições não-normais, têm-se apenas resultados assintóticos, isto é, a distribuição χ^2_{n-p} pode ser usada, somente, como uma aproximação, que em muitos casos pode ser pobre. Além disso, χ^2 tem como desvantagem o fato de tratar os y_i's simetricamente. Em muitos casos, é preferida em relação à *deviance*, por facilidade de interpretação (Demétrio, 2002).

Tanto a *deviance* quanto a estatística χ^2 generalizada de Pearson tem uma distribuição χ^2 exata para os modelos lineares normais. Porém, a *deviance* tem uma vantagem geral como medida de discrepância por ser aditiva para conjuntos de modelos aninhados se as estimativas de máxima verossimilhança forem utilizadas. Contudo, a estatística χ^2 as vezes pode ser preferida porque é uma interpretação mais direta (McCullagh and Nelder, 1989; Turkman and Silva, 2003).

3.5. Estimação do parâmetro ϕ

No caso das distribuições binomial e Poisson, $\phi = 1$. ϕ é desconhecido nas distribuições normal e normal inversa ($\phi = \sigma^2$) e gamma ($\phi = \nu^{-1}$) e admite-se que seja constante para todas as observações. Os métodos mais usados para a estimação de ϕ são: método da máxima verossimilhança, método dos momentos e perfil de verossimilhança (Demétrio, 2002).

O método da máxima verossimilhança é sempre possível em teoria, mas pode-se tornar intratável computacionalmente quando não existe solução explícita. Se ϕ é o mesmo para todas as observações, a estimativa de máxima verossimilhança de β independe de ϕ mas já a matriz de variâncias e covariâncias dos β's envolve esse parâmetro. Interpretando o logaritmo da função de verossimilhança $l(\boldsymbol{\beta}, \phi; \mathbf{y})$como função de β e de ϕ, **y**, a estimativa de máxima verossimilhança para ϕ é obtida, fazendo-se:

$$\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \phi} = 0 .$$

Para as distribuições normal e normal inversa tem-se

$$\hat{\phi} = \frac{1}{n} D^* \ .$$

3.6. Testes de hipóteses

Nos MLGs os métodos de inferência baseiam-se, fundamentalmente, na máxima verossimilhança. De acordo com esta teoria, existem três estatísticas para testar hipóteses relativas aos parâmetros β's (Demétrio, 2002). São elas:

- Teste da Razão de verossimilhança ou Estatística de Wilks;
- Teste de Wald; e
- Teste escore ou Estatística de Rao.

Estas estatísticas são assintoticamente equivalentes e, sob H_0 e para ϕ conhecido, convergem para uma variável com distribuição χ^2_p, sendo. A razão de verossimilhança é o critério que define um teste mais poderoso (Demétrio, 2002).

3.6.1. Teste da razão de verossimilhança

Envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição $\left(\ell\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2; \mathbf{y}\right)\right)$ e sob H_0 $\left(\ell\left(\boldsymbol{\beta}_{1,0}, \hat{\boldsymbol{\beta}}_{2,0}; \mathbf{y}\right)\right)$, ou, em termos de *deviance*, a comparação de $D(\mathbf{y};\hat{\boldsymbol{\mu}})$ e $D\left(\mathbf{y};\hat{\boldsymbol{\mu}}_0\right)$ em que $\hat{\boldsymbol{\mu}}_0 = g^{-1}\left(\hat{\boldsymbol{\eta}}_0\right)$ e $\hat{\boldsymbol{\eta}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}}_0$. Esse teste é, geralmente, preferido no caso de hipóteses relativas a vários coeficientes β's. Se as diferenças são grandes, então, H_0 é rejeitada. A estatística para esse teste é dada por:

$$\Lambda = -2\ln\lambda = 2\left[\ell\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2; \mathbf{y}\right) - \ell\left(\boldsymbol{\beta}_{1,0}, \hat{\boldsymbol{\beta}}_{2,0}; \mathbf{y}\right)\right] = \frac{1}{\phi}\left[D\left(\mathbf{y};\hat{\boldsymbol{\mu}}_0\right) - D\left(\mathbf{y};\hat{\boldsymbol{\mu}}\right)\right] \ .$$

Para amostras grandes, rejeita-se H_0, a um nível de probabilidade com α100%, se $\Lambda \sim \chi^2_{q,1-\alpha}$ (Turkman e Silva, 2000; Demétrio, 2002).

3.6.2. Teste de Wald

É baseado na distribuição normal de $\hat{\boldsymbol{\beta}}$ e é uma generalização da estatística *t* de Student (Wald, 1943). É, geralmente, o mais usado no caso de hipóteses relativas a um único coeficiente β_j. As funções estimáveis de β podem ser utilizadas para fazer a inferência nos MLGs. O resultado básico é

$$\mathrm{Var}\left[\hat{\boldsymbol{\beta}}\right] = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

Segue-se que

- A variância de uma função estimável $\mathbf{K}'\hat{\boldsymbol{\beta}}$ é

$$\mathrm{Var}\left[\mathbf{K}'\hat{\boldsymbol{\beta}}\right] = \mathbf{K}'\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{K}$$

- A estatística de Wald para $H_0 : \mathbf{K}'\hat{\boldsymbol{\beta}} = \mathbf{K}'\hat{\boldsymbol{\beta}}_0$ é

$$\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{K}'\hat{\boldsymbol{\beta}}_0\right)'\left[\mathbf{K}'\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{K}\right]^{-1}\left(\mathbf{K}'\hat{\boldsymbol{\beta}} - \mathbf{K}'\hat{\boldsymbol{\beta}}_0\right)$$

Retomando $\mathbf{W} = \mathbf{D}'\left(\mathbf{R}_\mu^{½}\mathbf{A}\mathbf{R}_\mu^{½}\right)^{-1}\mathbf{D}$. Se **A** é conhecido, como no caso das distribuições Poisson e binomial, por exemplo, então, a estatística de Wald tem distribuição aproximada de χ_ν^2, onde ν = rank(**K**).

Testes F

Quando **A** depende de um parâmetro de escala ϕ desconhecido (como no caso da distribuição normal), então **W** depende das estimativas de **A**. Nestas situações, a estatística de Wald dividida pelo rank(**K**) tem distribuição aproximada $F_{(\nu 1,\nu 2)}$, onde ν_1 = rank(**K**) e ν_2 = são os graus de liberdade associados com a estimativa de ϕ. a razão de Wald e o rank(**K**) podem ser utilizadas para testar H_0: $\mathbf{K}'\widehat{\boldsymbol{\beta}} = 0$.

3.6.3. Teste escore

Este teste tem sido muito usado na Bioestatística. A estatística de teste é dada por:

$$E = \mathbf{U}_1'\left(\hat{\boldsymbol{\beta}}_0\right)\ \hat{\mathrm{Var}}_0\left(\hat{\boldsymbol{\beta}}_1\right)\mathbf{U}_1\left(\hat{\boldsymbol{\beta}}_0\right),$$

sendo $\hat{\mathrm{Var}}_0\left(\hat{\boldsymbol{\beta}}_1\right)$ a $\mathrm{Var}_0\left(\hat{\boldsymbol{\beta}}_1\right)$ avaliada em $\hat{\boldsymbol{\beta}}_0 = [\boldsymbol{\beta}_{1,0}' \quad \hat{\boldsymbol{\beta}}_{2,0}']'$. Para amostras grandes, rejeita-se H_0, a um nível de probabilidade com $\alpha 100\%$, se $E \sim \chi^2_{q,1-\alpha}$ (Demétrio, 2002).

3.6.4. Caso particular:

No caso em que há interesse no teste de hipótese do vetor $\boldsymbol{\beta}$ como um todo

$$\mathrm{H}_0: \beta = \beta_0 \ \textit{versus}\ \mathrm{H}_a: \beta \neq \beta_0$$

o vetor $\boldsymbol{\beta}_2$ desaparece e $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$ (q passa a ser igual a p), e têm-se as expressões:

a) Teste da razão de verossimilhanças:

$$\Lambda = -2\ln\lambda = 2\left[\ell\left(\hat{\boldsymbol{\beta}};\mathbf{y}\right) - \ell\left(\boldsymbol{\beta}_0;\mathbf{y}\right)\right].$$

b) Teste de Wald:

$$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)'\hat{\mathfrak{I}}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right),$$

sendo $\hat{\mathfrak{I}}$ a matriz de informação de Fisher avaliada em $\hat{\boldsymbol{\beta}}$;

c) Teste escore:

$$E = \mathbf{U}'\left(\boldsymbol{\beta}_0\right)\ \mathfrak{I}_0^{-1}\left(\hat{\boldsymbol{\beta}}_1\right)\mathbf{U}\left(\boldsymbol{\beta}_0\right),$$

sendo $\mathfrak{I}_0$ a matriz de informação de Fisher avaliada em $\boldsymbol{\beta}_0$ (Demétrio, 2002).

3.7. Intervalos de confiança

Pode-se utilizar qualquer uma das três estatísticas de teste supracitadas para se construir os intervalos de confiança para $\boldsymbol{\beta}_1$. A partir da estatística de teste da razão de verossimilhanças, uma região de confiança para $\boldsymbol{\beta}_1$, com um coeficiente de confiança de $100(1 - \alpha)\%$, inclui todos os valores de $\boldsymbol{\beta}_1$ tais que:

$$2\left[\ell\left(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2; \mathbf{y}\right) - \ell\left(\boldsymbol{\beta}_{1,0}, \hat{\boldsymbol{\beta}}_{2,0}; \mathbf{y}\right)\right] < \chi^2_{q,1-\alpha},$$

sendo $\hat{\boldsymbol{\beta}}_{2,1}$ a estimativa de máxima verossimilhança de $\boldsymbol{\beta}_2$ para cada valor de $\boldsymbol{\beta}_1$ que é testado ser pertencente, ou não, ao intervalo (Demétrio, 2002).

Para a estatística do teste de Wald, uma região de confiança para $\boldsymbol{\beta}_1$, com um coeficiente de confiança de $100(1 - \alpha)\%$, inclui todos os valores de $\boldsymbol{\beta}_1$ tais que:

$$\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right)' \left[\hat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_1\right)\right] \left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) < \chi^2_{q,1-\alpha}$$ (Demétrio, 2002).

3.8. Técnicas para a verificação do ajuste

Demétrio (2002) e Turkman e Silva (2000) destacam que a escolha adequada de um MLG engloba três passos:

1) A definição da distribuição;
2) A definição da função de ligação; e
3) A definição da matriz do modelo.

Mesmo após a escolha cuidadosa de um modelo e um ajuste aos dados, pode acontecer do resultado obtido não ser satisfatório. Podendo ser em decorrência de algum desvio sistemático entre valores observados e ajustados devido a uma escolha inadequada de um dos três passos da escolha de um MLG ou, por existir um valor discrepante em relação aos demais valores. Mas, na prática, em geral, há uma interação dos diferentes tipos de falhas (Demétrio, 2002).

3.9. Aplicações

Os MLGs são uma ferramenta de fundamental importância na análise estatística, porém ainda não é uma técnica muito utilizada no contexto do melhoramento animal. Contudo, sua aplicação cabe aos mais diversos campos de estudo como por exemplo, no caso de (Rocha et al., 2014) aplicaram a metodologia dos MLGs com a finalidade de estudar a relação entre a aplicação de silício, potássio e cálcio em coentro e o número de estômatos nas faces abaxial e adaxial das folhas do coentro, no presente estudo foi considerada a distribuição Poisson, com função de ligação a função logarítmica. O modelo considerado ajustou-se bem aos números de estômatos na folha de coentro. A aplicação de Silício e relações K:Ca não teve efeito sobre o número de estômatos na face abaxial das folhas de coentro, porém para o número de estômatos na face adaxial, o modelo com a interação destes fatores deve ser considerado.

Em um outro contexto, Rios-Neto e Oliveira (1999) utilizaram a metodologia dos MLGs para o desenvolvimento de modelo log-lineares que servissem de base para as análises de tendência de participação na população economicamente ativa e racionalizar a teoria de projeção.

E ainda temos o exemplo de Venezuela (2003) que objetivou com seu estudo o desenvolvimento de um sistema alternativo para a obtenção de equações para os MLGs, considerando dados com medidas repetidas, bem como obter as estimativas do parâmetros por meio de processos iterativos.

4. MODELOS LINEARES GENERALIZADOS MISTOS

O Modelo Linear Generalizado Misto (MLGM) é uma extensão do MLG e do MLM para acomodar distribuições não-normais com efeitos aleatórios normais. Sua formulação é relativamente simples (Littell et al., 2006). No caso dos MLGM, o preditor linear $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, passa a incluir os efeitos aleatórios, sendo

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}\,.$$

E o MLGM tem a formulação geral:

$$g\left(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}}\right) = \mathbf{Zu} + \mathbf{X\beta},$$

onde **Y** é a variável resposta; g(.) a função de ligação que é linear nas variáveis explanatórias; $\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}} = \mathrm{E}[\mathbf{Y}|\mathbf{U} = \mathbf{u}]$ é a esperança da resposta condicional aos efeitos aleatórios; **β** é o vetor de efeitos fixos; **X** é a matriz do modelo que relaciona os efeitos fixos a $g(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}})$; **u** é um vetor de efeitos aleatórios. Cada fator de agrupamento do efeito aleatório tem distribuição $\mathbf{u} \sim N(0, \mathbf{I}\sigma_u^2)$, onde $N(.,.)$representa a distribuição normal com média e variância indicada no parênteses; σ_u^2 componente de variância; **I** é uma matriz identidade; e **Z** é a matriz do modelo que relaciona $g(\boldsymbol{\mu}_{\mathbf{Y}|\mathbf{U}})$ a **u**. As estimativas dos parâmetros podem ser obtidas pela minimização da aproximação Laplaciana para a função de desvio (Vazquez et al., 2010).

A formação de um MLGM segue a mesma estratégia básica de um MLG, com foco nos momentos condicionais. Isto é, a forma da verossimilhança determina a **função de variância** e contém o parâmetro natural. A função da média θ(μ) pode ser usada como uma **ligação canônica**, ou qualquer outra função de μ pode ser utilizada. A dificuldade de ajustar os MLGMs não está na formulação matemática dos modelos, no cálculo da verossimilhança de **Y** como uma base da inferência estatística (Littell, et al., 2006).

Os MLGs estão baseados na **função de verossimilhança** dos dados. No Modelo Linear Misto (MLM), a estimação e a inferência também estão baseadas na função do logaritmo da verossimilhança **marginal** ou **residual** dos dados. A distribuição marginal é obtida através da integração da distribuição conjunta dos dados e os efeitos aleatórios sobre os efeitos aleatórios (Littell et al., 2006). Em um modelo Gaussiano, a matriz de variância e covariância da distribuição marginal dos dados é

$$V\left(\mathbf{y}\right) = \mathbf{Z}\left(\mathbf{I}\sigma_{\boldsymbol{u}}^2\right)\mathbf{Z}' + \mathbf{R},$$

onde $V(\mathbf{y})$ é a variância da variável resposta *y*; σ_u^2 é a variância do fator aleatório, cujos níveis são assumidos como sendo independentes e identicamente distribuídos; e **R** é a matriz de covariâncias dos resíduos, independentes e identicamente distribuídos

($\mathbf{R} = \mathbf{I}\sigma_u^2$). Em análises genéticas, a matriz de variâncias e covariâncias de efeito de touro ou efeito de animal, **u** (para um modelo touro ou modelo animal), tipicamente resulta em uma expressão para a distribuição marginal dos dados de

$$V(\mathbf{y}) = \mathbf{Z}\left(\mathbf{A}\sigma_{\boldsymbol{u}}^2\right)\mathbf{Z}' + \mathbf{R},$$

onde $\mathbf{A}\sigma_u^2$ é a matriz de covariância do vetor multivariado de efeitos aleatórios **u**, e **A** é a matriz de relacionamento aditivo. Animais são geneticamente relacionados a outros, então é esperado que seus desempenhos sejam correlacionados, a menos que σ_u^2 seja 0 (Vazquez et al., 2010).

Quando a distribuição de **Y**|**u** é *não* normal, a obtenção da distribuição marginal é geralmente difícil. De maneira alternativa, pode ser aplicada a estimação do modelo linear misto repetidamente em um modelo aproximado. Isto é a abordagem da **pseudo-verossimilhança** (Littell, et al., 2006).

4.1. Estimação dos parâmetros

4.1.1. Abordagem da Pseudo-verossimilhança

O ajuste de um MLG pode ser realizado através da solução de forma iterativa da equação de mínimos quadrados ponderados para um modelo linear

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{y}^*,$$

onde $\mathbf{y}^*$ é a variável dependente que está sendo estudada, também chamada de pseudo-dados. O processo é iterativo porque $\mathbf{y}^*$, assim como a matriz de ponderações **W**, dependem das estimativas atuais. Uma abordagem para motivar o método de ajuste é aplicar uma série de Taylor de primeira ordem para a média do modelo $g^{-1}(\mathbf{x}'\boldsymbol{\beta})$ para a média. A expansão depende do valor de $\boldsymbol{\beta}$ que é escolhido na linearização. O resultado é um modelo de regressão linear com variâncias desiguais que dependem de $\boldsymbol{\beta}^*$. Este modelo pode ser ajustado através de um *software* padrão que gera as estimativas de $\boldsymbol{\beta}$.

Com base nestas novas estimativas, $\mathbf{y}^*$ e os pesos das variâncias são atualizados e é derivado um novo modelo linear. Este processo continua até que os efeitos fixos não mudem mais (Littell, et al., 2006).

No caso dos MLGM, a mesma ideia é aplicada. Remove-se a não-linearidade aplicando uma expansão de Taylor de primeira ordem para $g^{-1}(\mathbf{X'\beta+Zu})$ sobre os valores atuais de $\boldsymbol{\beta}$ e $\mathbf{u}$ (Littell, et al., 2006).

Breslow e Clayton (1993) e Wolfinger e O'Connell (1993) demonstraram que as soluções para $\boldsymbol{\beta}$ e $\mathbf{u}$ podem ser obtidas através da solução iterativa das equações de modelos mistos generalizados:

$$\begin{bmatrix} \mathbf{X'WX} & \mathbf{X'WZ} \\ \mathbf{Z'WX} & \mathbf{Z'WZ+G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'Wy}^* \\ \mathbf{Z'Wy}^* \end{bmatrix},$$

onde $\mathbf{W}$ e $\mathbf{y}^*$ são definidos como nas equações de solução para o MLG apresentado anteriormente. Que é,

$$\mathbf{W} = \mathbf{D'R}^{-1}\mathbf{D}$$
$$\mathbf{y}^*\boldsymbol{\eta} = \hat{\boldsymbol{\eta}} + (\mathbf{y} - \boldsymbol{\mu})\mathbf{D}^{-1}$$
$$\mathbf{D} = [\partial\boldsymbol{\mu} / \partial\boldsymbol{\eta}]$$
$$\mathbf{R} = \left(\mathbf{R}_{\mu}^{1/2}\mathbf{A}\mathbf{R}_{\mu}^{1/2}\right) .$$

Os procedimentos de Breslow-Clayton e Wolfinger-O'Connell são similares no fato de que ambos utilizam equações de MLG. A principal diferença entre o termo de Quasi-Verossimilhança Penalizada (QVP) de Breslow e Clayton (1993) e dos termos de Pseudo-Verossimilhança (PV) ou Pseudo-Verossimilhança Restrita (PVR) de Wolfinger e O'Connell (1993) está na estimação do parâmetro ϕ. No procedimento de Breslow-Clayton o parâmetro de escala é fixo em $\phi = 1$, enquanto que no procedimento de Wolfinger-O'Connell, é sempre estimado (Littell, et al., 2006).

4.2. Funções Procedimento

Assim como em um MLC, a principal ferramenta de inferência nos MLGM é a função previsível $\mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$. A lógica que conduz os objetivos de um estudo particular para seleção de funções previsíveis que abordam esses objetivos é idêntica à de modelos mistos padrão. Sendo $\mathbf{L}'=[\mathbf{K}'\mathbf{M}]$, a variância do erro de predição de uma função previsível é

$$\mathrm{Var}\left[\mathbf{K}'\hat{\boldsymbol{\beta}} + \mathbf{M}'\left(\hat{\mathbf{u}} - \mathbf{u}\right)\right] = \mathbf{L}'\mathbf{C}\mathbf{L},$$

onde

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{W}\mathbf{X} & \mathbf{X}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{X} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-}$$

(Littell, et al., 2006).

4.3. Estatística de Wald

As hipóteses sobre as funções previsíveis podem ser testadas utilizando a estatística de Wald ou a estatística *F*. a fórmula básica da estatística de Wald é

$$F_w = \left(\mathbf{L}'\hat{\boldsymbol{\alpha}}\right)'\left(\mathbf{L}'\mathbf{C}\mathbf{L}\right)^{-1}\left(\mathbf{L}'\hat{\boldsymbol{\alpha}}\right),$$

onde $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}}\left(\hat{\mathbf{u}} - \mathbf{u}\right)$ (Littell, et al., 2006).

A estatística de Wald tem distribuição aproximada χ^2_v , onde v = rank(**L**). Quando a variância condicional **R**, depende de um parâmetro de escala (ou seja, **A** é conhecida), então o teste χ^2 pode ser utilizado (Littell, et al., 2006).

Testes F

Se **R** depender de um parâmetro de escala desconhecido, é preferido que se divida a estatística de Wald pelo rank(**L**) (Littell, et al., 2006).

$$F = F_w / \text{rank}(\mathbf{L})$$

tem distribuição aproximada $F_{(\nu_1,\nu_2)}$, onde ν_1 = rank(**L**) e ν_2 são os graus de liberdade usados para estimar **L′CL**. Em casos simples, ν_2 corresponde aos graus de liberdade necessários para a estimação de ϕ (ex.: σ^2 em um modelo com erros normais). Em casos mais complexos, por exemplo, tem que fazer uma aproximação em ν_2 utilizando um procedimento do tipo Satterhwaite (Littell, et al., 2006).

Tais procedimento é apropriado quando o parâmetro de escala ϕ é conhecido. Quando ϕ é desconhecido (em casos de erros normais) ou quando pressuposições nominais sobre ϕ são violadas, devem ser utilizados procedimentos capazes de estimar ϕ, tais como PV ou PVR (Littell, et al., 2006).

4.4. Função do parâmetro de escala

Para Littell et al. (2006) não há razão para se definir $\phi = 1$ para distribuições da família exponencial que contém parâmetro de escala livre (normal, normal inversa e gamma). Por outro lado, se a distribuição condicional não contiver um parâmetro livre (Bernoulli, binomial, Poisson etc). Não deve ser necessariamente adicionada à estimação por *default*.

Um parâmetro de escala multiplicativo na função de variância é um "dispositivo" comum para se considerar a superdispersão nos modelos generalizados. Por exemplo, ao se ajustar uma regressão Poisson com a função de ligação log e se determinar que há uma superdispersão nos dados, a função de variância de Poisson $V(\mu) = \exp\{\mathbf{x}'\boldsymbol{\beta}\}$ deve ser substituída por $\phi V(\mu)$. Dessa forma, o parâmetro de superdispersão ϕ "ajusta" para o fato de que os dados são mais variáveis do que se espera de dados sobre a pressuposição de Poisson (Littell, et al., 2006).

Existem importantes consequências quando se adiciona um parâmetro de escala à função de variância. Primeiro, a superdispersão como um fenômeno só possui um significado relativo para a distribuição de base. Dados com superdispersão em relação a distribuição Poisson podem não ser superdispersos em relação a distribuição binomial negativa. Segundo, é importante que as razões da superdispersão sejam investigadas, pois pode apontar uma importante quebra no modelo que não deve ser "remendada" ou encoberta pela função de variância escalonada (Littell, et al., 2006).

A superdispersão pode estar relacionada a:

- **Dados correlacionados**: associações positivas entre as observações criam uma situação onde o número de observações efetivas é reduzido em relação a um conjunto de dados independentes do mesmo tamanho. Isto é, os n dados correlacionados não fornecem a mesma quantidade de informação do que n observações independentes. A variabilidade de estatísticas, portanto, aumenta em comparação com o caso de independência. O curso de ação apropriado é considerar as correlações no modelo usando, por exemplo, efeitos aleatórios.
- **Variáveis omitidas**: a dispersão residual em um modelo aumenta se variáveis importantes são omitidas da análise. Como resultado, a estimativa padrão de ϕ com base nos resíduos aumenta. As ações apropriadas neste caso são: não multiplicar a variância pelo parâmetro escalar estimado a partir dos dados e corrigir o modelo.
- **Distribuição mal especificada**: os dados podem aparecer superdispersos porque estes podem são seguir a distribuição que foi assumida. Um exemplo típico é a inflação de zeros em processos de contagem. Muitos dados podem apresentar um excesso na quantidade de zeros em comparação a distribuição dita Poisson. Uma razão poderia estar no fato de que estes zeros poderiam ter sido gerados a partir de dois processos: um que produz zeros com probabilidade π e um processo de Poisson. Os dados resultantes são inflacionados em zero por uma mistura dos dois processos. Assumindo que somente o processo de Poisson e aumentando a variância proporcionalmente para todas as observações, não é necessária a correção (Littell, et al., 2006).

Gianola e Foulley (1983) apresentam um método não-linear para a avaliação de variáveis categóricas ordenadas, no contexto de um banco de dados frequentemente encontrado no melhoramento animal. O modelo assume uma variável contínua adjacente

que é descrita como uma combinação linear de variáveis amostradas de distribuições conceituais. Em contraste a outros métodos de análise de dados categóricos, esse procedimento, leva em consideração a pressuposição de que candidatos a seleção são amostrados de uma distribuição com média e variância conhecidas previamente. Problemas teóricos que surgem quando modelos lineares são aplicados a dados categóricos são eliminados pois o procedimento ajusta automaticamente as diferenças de incidência entre as populações consideradas na análise. Além disso, o método pode ser mais generalizado para levar em consideração os efeitos de variáveis concomitantes.

A metodologia apresentada por estes autores foi desenvolvida com base na metodologia elaborada por Thompson (1979), que sugeriu uma alternativa para considerar dados com distribuição binomial com valor médio $\Phi(\mathbf{X\beta + Zu})$. Neste conjunto, as estimativas de máxima verossimilhança de $\boldsymbol{\beta}$ e $\mathbf{u}$ poderiam ser obtidas de forma iterativa a partir de um conjunto de equações similares para Mínimos Quadrados Ponderados, com o vetor de observações $\mathbf{y}$ substituído por $\mathbf{X\beta + Zu + W [Y - \Phi(X\beta + Zu)]}$ onde $\mathbf{W}$ é diagonal e também com uma matriz diagonal substituindo a matriz de covariâncias residuais. Thompson denominou esta metodologia como Modelo Linear "Generalizado" no sentido de Nelder e Wedderburn (1972), no qual de $\boldsymbol{\beta}$ e $\mathbf{u}$ são considerados constantes. Se $\mathbf{u}$ for um vetor de valores reais de variáveis aleatórias ao invés de constantes, Thompson (1979) disse que seria intuitivamente interessante modificar este Modelo Linear "Generalizado" na mesma forma que dos mínimos quadrados ponderados são aperfeiçoados para obter as equações de modelos mistos.

Tempelman (1998) propôs uma extensão hierárquica para os MLGMs utilizados na análise genética de características de aptidão e fertilidade de gado de leite. Tal modelo permite que padrões de dispersão complexos que acomodem heterocedasticidade e robustez. Baseado na vantagem de que avanços estáveis no poder computacional disponível têm facilitado a análise multicaracterística envolvendo medidas contínuas e discretas. Ele ainda afirma que a inferência bayesiana via desenvolvimento dos métodos de Monte Carlo Cadeias de Markov (MCMC) permite cada vez mais generalidade e dimensões no modelo genético.

5. MODELOS DE EQUAÇÕES ESTRUTURAIS

Para elucidar a história da modelagem de equações estruturais, é interessante comentar a origem dos modelos que estão relacionados e a ordem cronológica de desenvolvimento: análise de regressão, análise fatorial, análise de trilha e os modelos de equações estruturais (Schumaker and Lomax, 2004).

O primeiro é o modelo de regressão linear, que usa um coeficiente de correlação e o critério de mínimos quadrados para calcular os pesos da regressão. Esta metodologia tornou-se possível graças à criação da fórmula para o coeficiente de correlação, por Karl Pearson em 1896, que fornece um índice para a relação entre duas variáveis. Um modelo de regressão consiste exclusivamente de variáveis observadas, onde uma única variável observada dependente (*y*) é predita ou explicada por uma ponderação linear, de uma ou mais variáveis observadas independentes (X's), que minimiza os valores da soma de quadrado residual (Schumacker e Lomax, 2004).

Alguns anos mais tarde, Charles Spearman (1904, 1927) utilizou o coeficiente de correlação para determinar quais os itens eram correlacionados para criar o modelo fatorial. Sua ideia básica era testar se um conjunto de variáveis era correlacionado, as respostas individuais do conjunto poderiam ser somadas para produzir um escore que mediria, definiria ou indicaria um constructo. Thurstone et al., em 1940 desenvolveram aplicações adicionais para os modelos de fatores e propuseram instrumentos que fornecem os valores observados dos constructos que poderiam ser inferidos. A análise de fatores foi utilizada em muitas disciplinas acadêmicas para a criação de instrumentos de mensuração por mais de 100 anos (Schumacker e Lomax, 2004).

Sewall Wright (1918, 1921, 1934), um biólogo, foi o responsável pelo desenvolvimento do terceiro tipo de modelo, a análise de trilha. Este modelo usa coeficientes de correlação e análise de regressão para modelar relacionamentos complexos entre variáveis observadas. As primeiras aplicações da análise de trilha lidavam com modelos de comportamento animal. Mas, infelizmente, foi negligenciada por muito tempo até que os econometristas se interessassem por essa técnica, nos anos 50, na forma de modelos de equações simultâneas, logo depois, os sociólogos redescobriram na década de 60. Em muitos aspectos, a análise de trilha envolve soluções para um conjunto de equações de regressões simultâneas, que teoricamente, estabelece o relacionamento entre variáveis observadas no modelo de trilha. Atualmente, sob o

contexto de equações estruturais, a análise de trilha é considerada um caso particular dos modelos de equações estruturais (MEEs) (Schumacker e Lomax, 2004).

O último tipo de modelo é o Modelo de Equação Estrutural. Os MEEs essencialmente combinam modelos de análise de trilha e modelos de na análise de fatores, ou seja, os MEEs incorporam variáveis latentes e observadas. O rápido desenvolvimento dos MEEs ocorreu devido a Karl Jöreskog (1973), Ward Keesling (1972) e David Wiley (1973); dessa forma, essa abordagem foi conhecida inicialmente como o Modelo JKW. Mais tarde ficou conhecida como Modelo de Relações Estruturais Lineares (Linear Structural Relations Model - LISREL) a partir do desenvolvimento do primeiro *software*, LISREL, em 1973 (Schumacker e Lomax, 2004).

A modelagem de equações estruturais vem se tornando uma escolha interessante dentre os métodos multivariados, e o periódico *Structural Equation Modeling* foi o recurso primário para o desenvolvimento de técnicas em modelagem de equações estruturais (Schumacker e Lomax, 2004).

5.1. Modelagem de Equações Estruturais

A modelagem de equações estruturais testa a plausibilidade de um modelo construído com base em uma teoria que sustenta o fenômeno estudado, da mesma maneira que outras técnicas de análise quantitativa multivariada e, se caracteriza pela capacidade de especificar, estimar e testar relações hipotéticas em um grupo de variáveis (Codes, 2005).

Essa modelagem permite considerar diversas relações possíveis entre variáveis, sendo elas dependentes ou independentes, ou seja, é possível analisar várias relações de dependência entre variáveis, incluindo a possibilidade de que uma variável dependente, em uma equação, seja independente em outra. Diferente dos procedimentos multivariados clássicos, que consideram múltiplas variáveis independentes (Codes, 2005).

A modelagem de equações estruturais utiliza vários tipos de modelos para descrever um relacionamento entre as variáveis observadas, com a finalidade de fornecer um teste quantitativo de um modelo teórico (Schumacker e Lomax, 2004). A possibilidade de que as variáveis sejam dispostas de modo intrincado dentro do modelo permite avaliar e estimar os efeitos diretos, indiretos e totais que uma variável pode exercer sobre outra variável (Codes, 2005). Isto é, vários modelos podem ser estudados

para testar a hipótese de como os conjuntos de variáveis definem as construções e como essas construções estão relacionadas entre si. Por exemplo, um pesquisador da área de educação pode supor que o ambiente doméstico de um aluno pode influenciar o seu sucesso na escola. Um pesquisador da área de marketing pode testar a hipótese de que o consumidor confia numa marca líder para aumentar as vendas de um produto desta companhia. Um profissional da saúde pode acreditar que uma dieta balanceada e exercícios regulares reduzem o risco de ataque cardíaco (Schumacker e Lomax, 2004).

Em cada exemplo supracitado, o pesquisador acredita que os conjuntos de variáveis definem as construções que têm suas hipóteses testadas para serem relacionadas de certa maneira, baseando-se na teoria e no conhecimento empírico. A meta da análise de um MEE é determinar a dimensão que é suportada pelo modelo teórico de acordo com os dados amostrais. Caso a amostra suporte o modelo teórico: modelos mais complexos podem ser testados; já quando a amostra não suporta o modelo teórico: o modelo original pode ser modificado e testado ou outros modelos precisarão ser desenvolvidos e testados (Schumacker e Lomax, 2004).

Em programas de melhoramento genético, o objetivo de seleção tende a envolver diversas características correlacionadas. Assim, os MEEs podem ser usados para representar o conjunto de características. Tais modelos constituem uma extensão do modelo multicaracterística padrão, através dos quais, a correlação entre características é tipicamente representada por associações lineares simétricas entre efeitos aleatórios considerados para cada característica, como efeito genético aditivo direto ou efeitos de ambiente permanente e temporário. Estas associações são representadas por componentes de covariância (Valente, 2010).

Diferentemente dos modelos multicaracterísticas (MMC), nos MEEs, uma característica pode ser uma função de outras características que pertencem ao conjunto de características estudadas, o que permite a representação de uma rede funcional entre elas. Os MEEs foram desenvolvidos com o objetivo de combinar informações qualitativas de causa e efeito com informações provenientes dos dados, a fim de fornecer uma estimativa quantitativa da relação de causa e efeito entre as variáveis de interesse (Valente, 2010).

Com MEE é possível estudar relações de recursividade (efeito de uma variável resposta em outra) e de *feedback* (ou simultaneidade) entre variáveis resposta. Com isso, uma relação complexa entre variáveis, que é muito comum em sistemas biológicos, pode ser representada de forma adequada por meio de associações lineares simétricas como

componentes de covariância em modelos multicaracterísticas clássicos (Gianola and Sorensen, 2004; Valente, 2010).

Os MEEs podem ser representados em sua forma geral como um sistema de equações em que cada equação é dada por:

$$y_j = f(\boldsymbol{y}_{pj}, e_j) \tag{20}$$

onde y_j é a variável dependente da equação, $\boldsymbol{y}_{pj}$ são variáveis dentre aquelas consideradas como "variáveis dependentes" do modelo, com índices diferentes de *j*, que influenciam y_j (denominados "pais" de y_j) e e_j representa o resíduo aleatório associado a y_j (Pearl, 2000; Valente, 2010). O modelo (20) é uma generalização não-linear e não-paramétrica de MEEs lineares:

$$y_j = \sum_{k \in pj} \lambda_{jk} y_k + e_j$$

em que *pj* compreende o conjunto de "variáveis dependentes" que são pais de y_j, e λ_{jk} é o coeficiente estrutural, e corresponde à modificação de valor esperada em y_j com respeito à variável y_k (Pearl, 2000).

Observa-se que para a representação de MEEs, é necessário definir *a priori*, para cada variável resposta *j* do conjunto estudado, quais das variáveis remanescentes serão consideradas como pais de *j*. Esta estrutura de associações causais pode ser representada por um gráfico contendo variáveis conectadas por setas. Este gráfico direcionado é denominado estrutura causal (Pearl, 2000). Considere como exemplo um MEE simples representando a associação entre três variáveis: y_1, y_2, y_3. Ao definir que y_1 influencia y_2 recursivamente, e que y_2 influencia y_3 de maneira semelhante, a estrutura causal recursiva acíclica descrita pode ser representada graficamente por $y_1 \rightarrow y_2 \rightarrow y_3$, e o sistema de equações pode ser representado por:

$$y_1 = e_1$$
$$y_2 = \lambda_{21} y_1 + e_2$$
$$y_3 = \lambda_{32} y_2 + e_3$$

Os MEEs sob o contexto de genética quantitativa foram apresentados por Gianola e Sorensen (2004). Segundo estes autores, um sistema de equações de duas características distintas, cujas observações do indivíduo *i* são representadas por y_{i1} e y_{i2}, pode ser descrito da seguinte forma:

$$y_{i1} = \lambda_{12} y_{i2} + \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + u_{i1} + e_{i1}$$
$$y_{i2} = \lambda_{21} y_{i1} + \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + u_{i2} + e_{i2}$$

onde $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ são vetores de efeitos fixos para a característica 1 e característica 2, $\mathbf{x}'_{i1}$ e $\mathbf{x}'_{i2}$ são vetores conhecidos de incidência dos efeitos fixos $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ nas observações, u_{i1} e u_{i2} são os efeitos genéticos aditivos e, e_{i1} e e_{i2} os resíduos do modelo. O parâmetro λ_{12} é a mudança no valor de y_{i1} em função do valor de y_{i2} e λ_{21} é a mudança no valor de y_{i2} em função do valor de y_{i1}. No caso descrito, existe um *feedback* ou um relacionamento simultâneo entre as características, pois λ_{12} e λ_{21} são diferentes de zero. Dessa forma, cada uma das características influencia de forma direta a outra característica e, de forma indireta a si própria. Já a recursividade ocorre quando apenas um dos dois coeficientes é diferente de zero. O modelo é descrito na forma de gráfico (Figura 1), por Valente (2010), no qual, os efeitos fixos são omitidos.

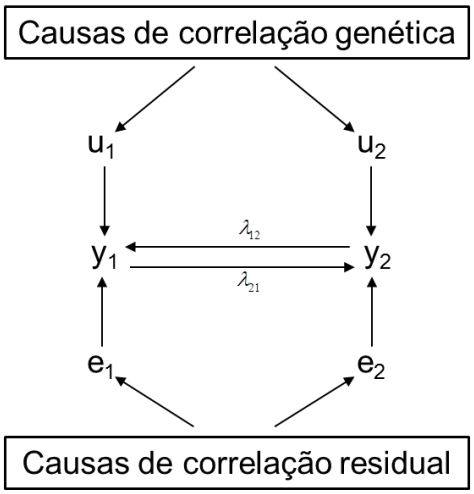


FIGURA 1 Modelo descrito por Valente (2010) que considera simultaneidade entre características: y_1 e y_2 são observações das características 1 e 2; u_1 e u_2 são efeitos genéticos aditivos e, e_1 e e_2 os resíduos atribuídos às características. Seta unidirecional

representa que a variável da ponta da seta é influenciada pela variável da base da seta. Setas opostas representam simultaneidade entre variáveis. O parâmetro λ_{ij} é a mudança na variável *i* relativa ao valor da variável *j*.

Considerando y_i como o vetor de observações de *t* diferentes características para o animal *i*, MEEs podem ser representados como:

$$\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i \tag{21}$$

onde $\mathbf{u}_i$ e $\mathbf{e}_i$ são vetores de valores genéticos aditivos e resíduos atribuídos a $\mathbf{y}_i$, $\boldsymbol{\beta}$ é o vetor que contém os efeitos fixos para as características, $\mathbf{X}_i$ é a matriz de incidência dos efeitos contidos em $\boldsymbol{\beta}$ no vetor $\mathbf{y}_i$ e $\mathbf{\Lambda}$ é uma matriz quadrada com ordem *t* que contém os coeficientes estruturais nas entradas fora da diagonal principal. A estrutura causal definida *a priori* indica quais entradas de $\mathbf{\Lambda}$ são consideradas como parâmetros livres e quais são obrigatoriamente iguais a zero (Valente, 2010).

A seguinte distribuição conjunta é considerada para $\mathbf{u}_i$ e $\mathbf{e}_i$:

$$\begin{bmatrix} \mathbf{u}_i \\ \mathbf{e}_i \end{bmatrix} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G}_0 & 0 \\ 0 & \mathbf{R}_0 \end{bmatrix} \right\},$$

em que $\mathbf{G}_0$ $\mathbf{G}_0$ e $\mathbf{R}_0$ são, respectivamente, matrizes de covariância genética aditiva direta e residual (Valente, 2010).

Com base no MEE (21), deriva-se o seguinte "modelo reduzido":

$$\begin{aligned} \left(\mathbf{I}_t - \mathbf{\Lambda}\right)\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i \\ \mathbf{y}_i &= \left(\mathbf{I}_t - \mathbf{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta} + \left(\mathbf{I}_t - \mathbf{\Lambda}\right)^{-1}\mathbf{u}_i + \left(\mathbf{I}_t - \mathbf{\Lambda}\right)^{-1}\mathbf{e}_i \end{aligned} \tag{22}$$

O modelo para *n* indivíduos é descrito por

$$\mathbf{y} = \left(\mathbf{\Lambda} \otimes \mathbf{I}_n\right)\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{23}$$

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0 \otimes \mathbf{A} & 0\\ 0 & \mathbf{R}_0 \otimes \mathbf{I}_n\end{bmatrix}\right\},$$

onde **y**, **u** e **e** são vetores de observações, efeitos genéticos aditivos e resíduos do modelo ordenados por característica e indivíduo dentro de característica, enquanto **X** e **Z** são matrizes de incidência dos efeitos em $\boldsymbol{\beta}$ e **u** no vetor **y** (Valente, 2010). O modelo (23) pode ser reescrito como:

$$\left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

de modo que o modelo reduzido se torna

$$\mathbf{y} = \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{Z}\mathbf{u} + \left[\mathbf{I}_{tn} - \left(\boldsymbol{\Lambda} \otimes \mathbf{I}_n\right)\right]^{-1}\mathbf{e}.$$

Considerando (22), observa-se que no modelo reduzido, o sistema é resolvido para as "variáveis resposta". Desta forma, o modelo resultante desta transformação corresponde ao modelo multicaracterísticas padrão:

$$\mathbf{y}_i = \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta} + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{u}_i + \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{e}_i$$
$$\mathbf{y}_i = \boldsymbol{\mu}_1^* + \mathbf{u}_i^* + \mathbf{e}_i^*$$

em que $\boldsymbol{\mu}_1^*$, $\mathbf{u}_i^*$ e $\mathbf{e}_i^*$ são respectivamente vetores de efeitos fixos, efeitos genéticos aditivos e resíduos de um modelo que não considera elos funcionais entre variáveis resposta. Adicionalmente, pode-se representar a distribuição conjunta dos efeitos aleatórios do modelo multicaracterísticas padrão como:

$$\begin{bmatrix}\mathbf{u}_i^*\\ \mathbf{e}_1^*\end{bmatrix} \sim N\left\{\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\mathbf{G}_0^* & 0\\ 0 & \mathbf{R}_0^*\end{bmatrix}\right\}$$
$$\mathbf{G}_0^* = \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1}\mathbf{G}_0\left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{'-1}$$

$$\mathbf{R}_0^* = \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{-1} \mathbf{R}_0 \left(\mathbf{I}_t - \boldsymbol{\Lambda}\right)^{'-1}$$

Desta forma, MEEs podem ser descritos como reparametrizações do modelo multicaracterísticas simples. As duas formas apresentadas são equivalentes, uma vez que geram a mesma distribuição para as variáveis resposta (Valente, 2010).

5.2. Estruturas Causais

A estrutura causal é a definição *a priori* de um subconjunto de variáveis remanescentes, para cada variável do sistema, que descreve de forma qualitativa os relacionamentos causais entre as características, através do efeito causal que uma característica exerce sobre a outra. Essa estrutura causal pode ser representada na forma de gráfico onde as variáveis (observadas ou não) constituem os "nós" e o relacionamento entre as variáveis é representado por setas direcionadas entre os nós. Ao se ajustar um MEE, é possível então inferir a magnitude de cada relacionamento causal pertencente à estrutura causal, que é quantificada pelos parâmetros do modelo, chamados de coeficientes estruturais (Rosa et al., 2011; Valente et al., 2011).

A análise de estruturas causais aborda os problemas de dependência entre variáveis, problemas estes típicos da regressão. Porém ela vai mais além, pois envolve duas ou mais equações de regressão na modelagem do fenômeno estudado. Isso faz com que tal método delineie problemas de regressão através de um "diagrama de trajetórias" (Codes, 2005).

Para Codes (2005) um aspecto importante dessa metodologia, está no fato de tais modelos serem lineares, uma vez que as relações entre todas as variáveis, podem ser representadas por equações lineares ou podem ser transformadas. Além disso, propicia a tradução das correlações entre as variáveis em um diagrama, possibilitando dessa forma, uma representação mais clara do fenômeno estudado. Essa técnica é usualmente aplicada à análise de fenômenos complexos e intrincados, assim, tais gráficos são dispositivos eficientes para ilustrar as relações simultâneas entre as variáveis.

Na genética quantitativa, tradicionalmente se estuda o relacionamento entre características via seu relacionamento probabilístico, através da utilização dos Modelos Multicaracterísticos (MMC) padrão. Mesmo que estes modelos sejam satisfatórios para

inferir os prováveis eventos, eles não são estáveis o suficiente para predizer o quanto as probabilidades mudariam a partir de intervenções externas (Rosa et al., 2011).

Informações a respeito de redes de fenótipos descrevendo cada inter-relação pode ser utilizada para predizer trajetórias subjacentes às características complexas tais como: doenças, crescimento e reprodução. Este fato pode ser utilizado para aperfeiçoar práticas de manejo e estratégias de seleção multicaracterística na produção animal (Rosa et al., 2011).

Em um MMC padrão os relacionamentos são representados por associações lineares simétricas entre as variáveis aleatórias. Um MEE, por sua vez, pode ser aplicado para estudar o relacionamento recursivo e/ou simultâneo entre as variáveis em um sistema multivariado. Logo, um MEE pode produzir uma interpretação do relacionamento entre características diferente dos MMC padrão. Além disso, em um MEE uma característica pode ser tratada como preditor para uma outra característica, indicando uma ligação funcional (causal) entre elas (Rosa et al., 2011).

Em seu estudo, Rosa et al. (2011) destacam ainda o fato de que um MEE pode ser usado no melhoramento animal sob a ótica da genética quantitativa clássica, mesmo quando não se tem informações de marcadores moleculares ou QTLs. Fato que bem ilustrado pela metodologia proposta por valente et al. (2010), pois esta metodologia permite a busca por estruturas causais recursivas, para a análise de características múltiplas no contexto de modelos mistos, mostrando que em certas condições, é possível inferir redes fenotípicas e efeitos causais compensando a ausência de informações de QTLs ou marcadores moleculares.

5.2.1. Terminologia

As estruturas causais podem ser representadas por diagramas constituídos por um conjunto de vértices (variáveis) conectados por linhas (*edges*) que representam as associações causais, quando possuem uma seta em uma das extremidades (linha direcionada ou *directed edge*) ou conexões diretas simétricas, quando não possuem seta em nenhuma das extremidades (linha não-direcionada ou *undirected ed*ge). Se todas as linhas de um gráfico forem direcionadas, este gráfico é considerado direcionado. Um Gráfico Acíclico Direcionado (GAD) é aquele representado por um conjunto de variáveis (nós ou *nodes*) conectados por setas direcionadas e que não possuem ciclos causais. Em

$A \rightarrow C \leftarrow B$, A e B (pais ou *parents*) são as causas diretas de C (filho ou *child*). Em $A \rightarrow B \rightarrow C$, A é causa indireta de C, mediada por B. Dois vértices são denominados adjacentes, quando estes são conectados por uma linha (Valente, 2010).

Uma trilha (*path*), em uma estrutura causal, é uma sequência de vértices conectados por linhas. Uma sequência de nós de uma trilha pode respeitar o sentido das setas, nesse caso temos uma trilha direta (*directed path*, ex.: $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$) ou não, então se trata de uma trilha indireta (*undirected path*, ex.: $A \rightarrow B \rightarrow C \leftarrow D \rightarrow E$). Em uma trilha indireta podem existir os chamados *colliders*, que são os vértices nos quais as setas apresentam uma convergência. Uma trilha é ativa quando é capaz de carregar informação ou dependência entre os vários vértices dos extremos. Desta forma, incondicionalmente toda trilha é ativa, a menos que tenha ao menos um *collider*, pois este bloqueia o fluxo de dependência. Exemplificando, A é marginalmente dependente de B em $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e $A \leftarrow C \rightarrow B$. Já em isso é $A \rightarrow C \leftarrow B$ ocorre. A capacidade de um vértice transmitir ou bloquear fluxos de dependência se inverte quando a trilha é analisada condicionalmente a ele. Como consequência, na estrutura $A \rightarrow C \leftarrow B$ condicionalmente a C, A e B são dependentes e independentes em $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e em $A \leftarrow C \rightarrow B$. A independência entre dois vértices pertencentes a um grupo de vértices remanescentes é chamada de d-separação (*d-separation*). Formalmente, A e B são considerados d-separados condicionalmente a um conjunto de vértices remanescentes **S** se não existir nenhuma trilha que permita o fluxo no GAD entre A e B, de modo que todos os *colliders* ou seus descendentes estejam em **S** e nenhum não-*collider* esteja em **S** (Rosa et al., 2011; Valente, 2010).

Sob algumas premissas, as d-separações em uma estrutura causal são refletidas como independências condicionais estatísticas na densidade conjunta das variáveis, o que é explorado na tentativa de selecionar estruturas causais a partir desta densidade conjunta. Deste modo, assume-se que um modelo causal impõe algumas marcas na densidade dos dados observados, e tenta-se recuperar a estrutura causal deste modelo a partir destas marcas. Esta tentativa assume uma conexão entre estrutura causal e distribuição de probabilidade. (Rosa et al., 2011; Valente, 2010).

5.2.2. Associação VS. Causalidade

Existe uma afirmação que diz: "correlação não implica causalidade". Esta afirmação é muito importante sob o ponto de vista da causalidade, ela enfatiza que o conhecimento da correlação entre duas variáveis não é suficiente para descrever a relação causal entre elas. A correlação entre duas variáveis *x* e *y* poderia ser oriunda de diferentes tipos de associações causais entre elas. Por exemplo, correlações entres estas variáveis poderiam aumentar devido a efeitos de causais de *x* em *y*, onde tal relacionamento casual é, geralmente representado por: $x \rightarrow y$, ou por um efeito causal de *y* em *x*, representado por $x \leftarrow y$, ou ainda a correlação pode ser devido a uma fonte comum de associação, ou seja, um conjunto de outras variáveis *z* que afeta tanto *x* quanto *y*, representado como $x \leftarrow z \rightarrow y$. Ou ainda, uma combinação desses três efeitos poderia estar explicando a correlação entre *x* e *y* (Rosa e Valente, 2012).

Conhecer a associação entre as variáveis é importante para entender como esta associação poderá ser utilizada na predição dos modelos para inferir como os eventos provavelmente são. No entanto, esta informação não é suficiente para predizer como intervenções externas mudariam as probabilidades (Pearl, 2000; Shipley, 2002; Pearl, 2009).

Em contrapartida, existe uma extensa literatura baseada em uma outra escola de pensamento, que afirma que há muito mais a ser aprendido a partir de dados observacionais do que simplesmente correlação e covariâncias entre variáveis (Spirtes et al., 2001; Shipley, 2002; Pearl, 2009). Segundo Tufte (2003), a não-equivalência entre correlação e causalidade negligencia informações sobre o relacionamento entre elas. Ele sugere que ao invés de "correlação não implica em causalidade", que usemos "a covariação observada empiricamente é uma condição necessária, mas não suficiente para a casualidade" ou "correlação não é causalidade, mas certamente é uma dica" (Rosa e Valente, 2012).

É importante que sejam realizados experimentos aleatorizados para tentar evitar que a associação observada entre as variáveis provenha de uma fonte extra de associação, oriunda de uma variável que não foi mensurada, que poderia estar afetando as variáveis estuadas. Entretanto, experimentos aleatorizados nem sempre são práticos e a não possibilidade de realizar tais experimentos complica o teste e a estimação dos efeitos causais. Além disso, uma desvantagem potencial para a realização de experimentos aleatorizados, está no fato de que as configurações experimentais na pesquisa pecuária,

em geral, não refletem as reais condições encontradas nas criações comerciais, sendo muito comum os efeitos inferidos a partir dos dados registrados em configurações experimentais representem uma superestimação dos mesmos efeitos quando testados em propriedades comerciais e, o mais importante é que o efeito de um fator específico é altamente dependente dos níveis de outros fatores (interação entre fatores), que podem ser variáveis em configurações comerciais, mas que foram mantidos constantes em experimentos controlados (Rosa e Valente, 2012).

5.2.3. Busca e recuperação das estruturas causais recursivas

Para ajustar um MEE é necessário a definição da estrutura causal, que consiste em definir *a priori*, para cada uma das variáveis resposta *i*, quais das variáveis remanescentes serão consideradas como pais de *i*. Uma estrutura causal recursiva pode ser representada por um GAD (Rosa et al., 2011). De acordo com a representação matricial de um MEE misto, a estrutura causal define quais entradas da matriz $\mathbf{\Lambda}$ serão consideradas como parâmetros livres em $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i + \mathbf{e}_i$ (Valente, 2010). A matriz $\mathbf{\Lambda}$ é a matriz dos coeficientes estruturais, coeficientes estes que definem a estrutura causal a ser especificada ao se ajustar o modelo (Rosa et al., 2011).

Na maioria das aplicações dos MEE em genética quantitativa, as estruturas causais são assumidas como conhecidas *a priori*. Entretanto, existem algoritmos desenvolvidos por pesquisadores das áreas de inteligência artificial e filosofia da matemática capazes de explorar espaços de hipóteses causais e buscar estruturas que são compatíveis com a distribuição conjunta apresentada pelas variáveis estudadas (Valente, 2010). Metodologias como o algoritmo IC (Iductive Causation) foram desenvolvidas para explorar a conexão entre as estruturas causais recursivas e as distribuições conjuntas e recuperam estruturas do GAD básico (ou classe de estruturas equivalentes) (Rosa et al., 2011).

O algoritmo IC torna possível a busca por estruturas causais recursivas compatíveis com a distribuição de probabilidade conjunta das variáveis consideradas, indicando que a aplicação de tais metodologias permite a seleção de estruturas causais sem confiar somente no conhecimento prévio. Mesmo assim, o algoritmo é construído com base em suposições relacionadas aos dados. Talvez a suposição mais forte se refira

à suficiência causal: assume-se que toda variável que influencia duas ou mais variáveis dentro do conjunto de variáveis estudadas já está dentro deste conjunto (Rosa et al., 2011).

Sob essa suposição, os resíduos do MEE para os quais a estrutura causal será escolhida serão considerados independentes entre as características. Esta construção é necessária para estabelecer a conexão entre as estruturas causais selecionadas e a distribuição de probabilidade conjunta no estudo, de forma que as d-separações nas estruturas causais entre as características sejam refletidas como correlações parciais nulas (Valente et al., 2011).

Neste cenário, o algoritmo IC considera uma matriz de correlação como *input* e busca por estruturas causais que sejam capazes de produzir aquela matriz, com suas dependências e independências condicionais. Porém, segundo Valente et al., (2010), fenótipos múltiplos podem apresentar efeitos genéticos correlacionados não observados que podem confundir a busca. Quando se usa um MEE de efeitos mistos para representar este cenário, este confundimento pode aparecer mesmo os resíduos sendo considerados como independentes. Como alternativa, Valente et al. (2010) propuseram uma metodologia com a qual se associa o ajuste de modelos Bayesianos e a aplicação do algoritmo IC à distribuição conjunta dos fenótipos condicionais aos efeitos genéticos. Com o propósito de validar e ilustrar seus métodos, Valente et al. (2010), fez uma aplicação em dados simulados com base em diferentes cenários (Valente et al., 2011).

Para promover a busca da estrutural causal, dentro do contexto de modelos mistos em genética quantitativa, Valente et al. (2010) adotou um ajuste de MEE com uma matriz de covariâncias residuais diagonal. Dentro desta construção, uma estrutura causal recursiva que é compatível com a distribuição conjunta de probabilidade dos dados pode ser buscada usando o algoritmo IC (Rosa et al., 2011).

No entanto, em um MEE misto com independência residual, associações entre características observadas são explicadas não somente pelas ligações causais entre elas, mas também por razões genéticas (Rosa et al., 2011).

5.2.4. Seleção da estrutura causal

Para ajustar MEEs a observações que pertencem a um conjunto de características, é necessário definir *a priori* a estrutura causal entre as variáveis estudadas. Considerando que o valor dos elementos fora da diagonal podem variar livremente ou serem definidos

como 0, existem potencialmente t (t-1) coeficientes estruturais e $2^{t(t-1)}$ estruturas que podem representar o relacionamento causal entre t características (Valente, 2010).

A seleção da estrutura causal a ser utilizada se torna um desafio para o ajuste de um MEE, devido ao aumento expressivo do número de possíveis estruturas na medida em que cresce o número de características estudadas, como demonstrado na TABELA 4. O grande número de hipóteses causais ocorre mesmo em situações restritas e com pequeno conjunto de características (Valente, 2010).

TABELA 4 – Número de possíveis estruturas causais que podem ser construídas dadas *n* variáveis resposta

n	Número de estruturas
2	4
3	64
4	4096
5	1048576
6	1073741824

(Shipley, 2002).

Shipley (2002) afirma que se fossemos testar uma estrutura em potencial por segundo, levaria quase 32 anos para testar cada estrutura potencial contendo 6 variáveis.

A comparação das estruturas utilizando critérios como o AIC (Akaike, 1976) ou o BIC (Schwartz, 1978) é exaustiva e considerada impossível. Nas aplicações recentes dos MEEs, no contexto de modelo misto em genética quantitativa, foram utilizadas crenças *a priori* a respeito da estrutura causal do sistema estudado, para que fossem selecionadas uma estrutura ou um pequeno grupo de estruturas, estas comparadas por critérios semelhantes aos supracitados (Valente, 2010).

Como demonstrado, o modelo reduzido (22) é equivalente ao MEE (23), uma vez que ambos produzem a mesma densidade de probabilidade:

$$N\left(\left(\mathbf{I}_t-\boldsymbol{\Lambda}\right)^{-1}\mathbf{X}_i\boldsymbol{\beta}+\left(\mathbf{I}_t-\boldsymbol{\Lambda}\right)^{-1}\mathbf{u}_i,\left(\mathbf{I}_t-\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Psi}_0\left(\mathbf{I}_t-\boldsymbol{\Lambda}\right)^{\prime-1}\right)=N\left(\boldsymbol{\mu}_t^*+\mathbf{u}_i^*,\mathbf{R}_0^*\right). \quad (24)$$

Mas, é evidente que os dois modelos não apresentam uma correspondência entre os seus parâmetros do tipo "um a um". Os MEEs apresentam coeficientes estruturais (presentes na matriz $\mathbf{\Lambda}$), além dos parâmetros de locação e de dispersão análogos àqueles do modelo reduzido (Valente, 2010).

O modelo multicaracterística padrão é identificável, uma vez que modificações nos valores dos parâmetros resultariam necessariamente em modificações na densidade de probabilidade dos dados por ele gerados. Como consequência, a inferência com base na função de verossimilhança torna-se possível para todos os parâmetros do modelo. Entretanto, a existência de parâmetros adicionais nos MEEs faz com que este modelo seja sub-identificável na função de verossimilhança, uma vez que mais de uma combinação de valores de parâmetros resultam em uma mesma densidade de probabilidade (24) é válida para infinitas combinações de valores de parâmetros no lado esquerdo da igualdade (Valente, 2010).

Como consequência da sub-identificabilidade do MEE apresentado, torna-se necessário aplicar restrições aos parâmetros do modelo para realizar inferências a respeito destes. Como exemplo, uma restrição suficiente para um modelo recursivo acíclico é considerar as covariâncias residuais iguais a 0 (Varona et al., 2007).

5.2.5. Inferência dos efeitos causais

Inferir os efeitos causais a partir de estudos de observação sempre requer suposições adicionais relativas a inferências estatísticas padrões que não incluem o significado causal (estimar correlações, coeficientes de regressão). No entanto, a informação fornecida por esses dois tipos de inferência é muito diferente. A informação estatística descreve essencialmente o quão plausível é um determinado evento. Por sua vez, a informação causal descreve como o valor de uma variável é afetado pelo valor de outras variáveis. A implicação prática é que a informação causal relacionada a um conjunto de variáveis permite que os resultados de intervenções externas na rede causal sejam preditos, o que pode ser muito útil para o manejo da produção animal (Rosa e Valente 2012).

O ajuste dos MEEs permite inferir os efeitos causais condicionais a uma estrutura causal dada *a priori*. Este tipo de análise permite predizer os resultados de intervenções externas a partir de dados observacionais e algumas pressuposições representadas pela

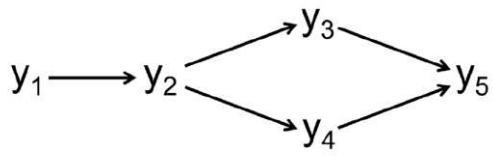
estrutura causal do modelo. Neste caso, a rede causal é representada por um gráfico direcionado, e cada família de nós envolvendo os pais e os filhos no gráfico representa uma equação estrutural onde a variável do lado esquerdo (Left-Hand Side - LHS) é determinada por uma função, geralmente construída como linear, das variáveis do lado direito (Right-Hand Side - RHS). Portanto, o sinal de igualdade nestas equações representa ao relacionamento assimétrico definido como 'é determinado por', que é diferente do significado usual das equações padrão. A interpretação é feita da seguinte maneira: se as variáveis na RHS assumissem alguns valores arbitrários específicos, a quantidade na LHS seria definida por uma função de variáveis no RHS. O significado causal das equações estruturais é essencialmente o mesmo que foi apresentado na formulação de Rubin (Rubin, 1974; Rosenbaum e Rubin, 1983), embora esses autores tenham articulado esse relacionamento em uma maneira diferente, usando conceitos como resultados potenciais e mecanismos de atribuição de tratamento (Rosa e Valente, 2012).

A interpretação causal derivada do ajuste de um MEE depende não só de pressuposições estatísticas, mas também de pressuposições causais. Por exemplo, inferências relacionadas com associações obtidas pelo ajuste de um MEE requerem assumir que a estrutura causal entre as variáveis é acíclica e que todas as variáveis têm efeitos causais sobre duas ou mais variáveis, com a construção da matriz de covariância residual como diagonal, o que é suficiente para garantir que qualquer MEE recursivo seja identificável a partir dos dados. No entanto, a pressuposição não é necessária para os MEEs, que apresentem ciclos e causa para associações devido às variáveis escondidas, representados pelas covariâncias residuais. Mesmo assim, a capacidade de identificação não é garantida se essas características são permitidas e, portanto, devem ser verificadas (Rosa e Valente, 2012).

A habilidade de predizer o efeito de intervenções é uma das mais importantes características fornecidas pela informação causal. Diferentes tipos de intervenções podem ser preditas a partir de modelos causais. O tipo mais comum de intervenção é a que configura externamente o valor da variável na rede. O efeito da intervenção pode ser representado pela eliminação de equações e também pela manipulação de modelos. Valente (2010) apresenta o modelo como exemplo:

$$\begin{cases} y_{i1} = \mu_1 + e_{i1} \\ y_{i2} = \mu_2 + \lambda_{21} y_{i1} + e_{i2} \\ y_{i3} = \mu_3 + \lambda_{32} y_{i2} + e_{i3} \\ y_{i4} = \mu_4 + \lambda_{42} y_{i2} + e_{i4} \\ y_{i5} = \mu_5 + \lambda_{53} y_{i3} + \lambda_{54} y_{i4} + e_{i5}, \end{cases}$$

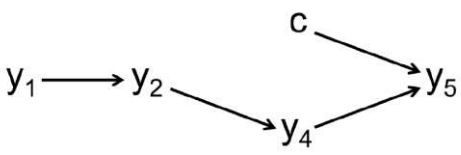
que representa uma estrutura descrita pela estrutura causal abaixo.



Ao inferir os efeitos da manipulação externa, através da eliminação de equações, é possível inferir o resultado dessa ação sem dados registrados sob manipulação. Por exemplo, o efeito da configuração de y_3 para uma constante arbitrária c seria representada por:

$$\begin{cases} y_{i1} = \mu_1 + e_{i1} \\ y_{i2} = \mu_2 + \lambda_{21} y_{i1} + e_{i2} \\ y_{i4} = \mu_4 + \lambda_{42} y_{i2} + e_{i4} \\ y_{i5} = \mu_5 + \lambda_{53} c + \lambda_{54} y_{i4} + e_{i5}, \end{cases}$$

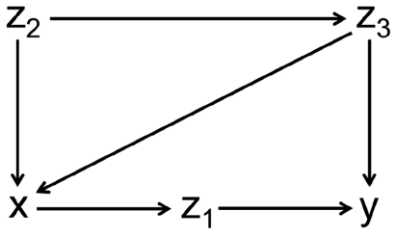
que pode ser representada graficamente pela estrutura causal a seguir. Aqui notamos que embora y_3 é fortemente conectado a y_2 e y_5, para valores diferentes, teriam consequências no valor de y_5, enquanto o valor de y_2 permaneceria inalterado, como descrito pelas equações estruturais. Que y_5, é definida como uma função de c, enquanto y_2 não é. Os valores de y_1 e y_4 também não são afetados pelas intervenções em y_3 (Rosa e Valente, 2012).



Para se inferir os efeitos das intervenções no MEE é necessário o conhecimento dos relacionamentos causais entre as características. Em casos onde as ligações entre as variáveis não são assumidas como sendo conhecidas, pode se utilizar algoritmos que permitem explorar o espaço das estruturas causais, tais como o algoritmo IC, o algoritmo SGS , o algoritmo PC entre outros. Estes algoritmos são baseados em uma série de suposições causais (Sprites et al., 2000), a partir das quais a suposição da suficiência causal é aparentemente a mais forte. Apesar disso, é possível obter algum aprendizado, mesmo se não assumirmos suficiência causal. Por exemplo, o algoritmo FCI (Fast Causal Algorithm) (Spirtes et al., 2000) perde a suficiência causal, embora resulte um *output* de um gráfico mais complexo. Informações prévias sobre as variáveis envolvidas (informação temporal ou associações causais já obtidas de experimentos aleatórios) pode auxiliar o aprendizado de associações causais com menos suposições e consequentemente melhorar as inferências num todo (Valente et al, 2011).

A representação da rede causal através de gráficos direcionados permite uma expressão eficiente das suposições subjacentes em tais redes. As independências condicionais probabilísticas esperadas que seguem a partir das suposições causais são dadas por d-separações (Pearl, 2000, 2009) no gráfico. O gráfico direcionado permite verificar se as suposições são suficientes para estimar o número alvo (como por exemplo, um único efeito causal).

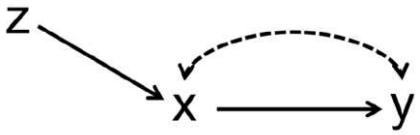
Como um exemplo desta aplicação, considere a possibilidade que a inferência dos efeitos causais entre duas características, x e y, é impedida pelo viés de confusão gerado a partir de trilhas envolvendo variáveis não mesuradas que são recursos adicionais de correlação entre elas, como as trilhas $x \leftarrow z_3 \rightarrow y$ e $x \leftarrow z_2 \rightarrow z_3 \rightarrow y$, como ilustrado na estrutura causal a seguir. Rosenbaum e Rubin (1983), usando o conceito de resultados potenciais para estudar modelos causais, formularam regras para se declarar um conjunto de variáveis concomitantes **Z** como suficiente para ser considerada na análise para permitir a identificação do efeito causal alvo de dados observacionais. Pearl (1995) formulou estas regras em termos de testes gráficos, chegando como o assim chamado critério *back-door*: o efeito causal total de x em y em um diagrama causal **G** pode ser considerado a partir de um conjunto de características mensuradas **S** (adicional à observação de x e y) tal que: 1) não-membro de **S** é um descendente de x e; 2) **S** d-separados de x e y no sub-gráfico formados pela exclusão em **G** de todas as setas emanando de x.



Na estrutura causal acima, para inferir o efeito causal total de x em y, seria suficiente para estudar as associações entre os dois nós condicionalmente em $\mathbf{S} = \{z_3\}$, que todas as trilhas *back-door*. Poderiam ser feitas inferências sobre este relacionamento causal (representado pelo coeficiente λ) pelo ajuste do seguinte modelo de regressão:

$$y = \lambda x + \beta z_3 + e.$$

Uma outra estratégia para inferir efeitos causais entre duas variáveis se dá pelo estudo de uma rede de associações mais ampla as contendo. Para o ajuste clássico, como representado a seguir, o coeficiente que representa o efeito causal alvo λ é determinado a partir da associação entre z e $x(\beta)$ e z e $y(\beta\lambda)$, sendo ambas identificáveis, então, $\lambda = \frac{\beta\lambda}{\beta}$. Esta estratégia de inferir efeitos causais não funciona, porém, para modelos causais com funções não lineares (ou não-específicas).

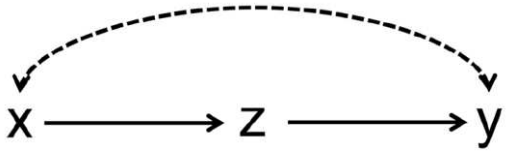


Inversamente, os relacionamentos causais estudados na base do critério *back-door* não precisam assumir nenhuma descrição dos relacionamentos causais. Como sempre, suposições casais são necessárias para estimar o efeito estimado sobre y originados da manipulação de x. O método de verificação combina a informação combina a informação no diagrama (com as trilhas *back-door* representadas por arcos bidirecionados ou variáveis latentes) e o uso de um conjunto de regras de inferência dado por Pearl (1995). Isto é desempenhado sem nenhuma suposição em relação à forma da função causal entre

características. Ao realizar a dedução baseadas nessas regras, pode se transformar o efeito causal procurado a partir de intervenções dentro de uma expressão equivalente, exclusivamente, envolvendo funções de probabilidade padrão. Por exemplo, considere que se está interessado em estudar o efeito causal da variável x sobre a variável y, aqui representada como $p(y \mid \tilde{x})$, onde $\tilde{x}$ significa que x é um conjunto de manipulação externa sobre a estrutura, que é ilustrada a seguir. Dentro dessa estrutura, existe uma trilha envolvendo variáveis não observadas responsáveis por algum grau de viés de confundimento representado por um arco bidirecionado, e o efeito causal alvo é mediado por z (Rosa e Valente, 2012). Pode ser provado que a informação alvo pode ser reduzida a:

$$p(y \mid \tilde{x}) = \sum_{x} p(z \mid x) \sum_{x'} p(y \mid x', z)\, p(x'),$$

que não apresenta variáveis sob intervenção, e encontra o critério *front-door* para identificar os efeitos causais (Pearl, 1995).



Note que a aplicação das regras já mencionadas pode tornar explícitos alguns aspectos que contradizem as práticas usuais em estudos similares. Para o exemplo dado, note que para identificar os efeitos causais de x em y, é necessário relatar uma variável z que mediu esse efeito, que geralmente é estimado como algo perdido, nesse caso. Embora contradizendo uma prática comum, o relacionamento causal obtido nesse ajuste depende de menos suposições que, por exemplo, o estudo dos efeitos causais via variáveis instrumentais, como na estrutura anterior. Utilizar variáveis instrumentais traz efeito casual identificado não somente a partir de suposições causais expressas pelo gráfico, como também a partir de linearidade assumida para os efeitos causais (Pearl, 2009). Todavia, pelo uso de cálculo para a intervenção, podemos verificar que o ajuste pode não permitir a identificação, se houver repressão a partir dos efeitos causais representados por funções paramétricas. Por outro lado, a linearidade não é assumida por inferência causal baseada no ajuste.

5.2.6. Propriedades de Markov

Cada estrutura causal carrega em si um conjunto de dependências e independências condicionais entre os vértices que a constituem. É possível observar que qualquer vértice do gráfico, condicional a seus pais, é independente de todos os vértices que não são seus descendentes. Adicionalmente, uma densidade conjunta é considerada compatível com uma estrutura causal gráfica se a primeira pode ser decomposta como em uma rede Bayesiana, na forma de um produto envolvendo a probabilidade de cada variável condicional aos seus pais de acordo com a estrutura causal (Compatibilidade de Markov). Nesse caso, para um conjunto de variáveis **V**, em que **pais**(*V*) é o conjunto dos pais de cada variável *V* em **V** (Spirtes et al., 2001):

$$p(\mathbf{V}) = \prod_{V \in \mathbf{V}} p(V \mid \mathbf{pais}(V))$$

A compatibilidade de Markov indica se uma estrutura causal é capaz de gerar determinada distribuição de probabilidade, por exemplo, a estrutura A_C_B não é capaz de gerar uma distribuição, na qual A é dependente de B condicional a C. Uma maneira de caracterizar o conjunto de distribuições compatíveis com um GAD é observar o conjunto de independências condicionais que a distribuição deve satisfazer. Com base no GAD, estas são obtidas no gráfico que representa a estrutura causal pelo critério da d-separação (Pearl, 2000).

Como fora mencionado, métodos de busca por estruturas causais têm base na conexão entre estas estruturas e densidades de probabilidade. Contudo, uma condição é necessária para que esta conexão seja denominada Condição Causal de Markov, é preciso que o modelo causal induza uma distribuição que satisfaça a compatibilidade de Markov, isto é, cada variável na distribuição deve ser independente das variáveis não descendentes, condicionalmente aos pais na estrutura causal. Esta condição é consequência de duas premissas: a) o compromisso de incluir no modelo todas as causas de duas ou mais variáveis respostas estudadas (suficiência causal); e b) a premissa de que não há associação entre pares de variáveis sem causalidade; isto é, uma variável causa à outra, ou elas apresentam uma causa comum (*Reichenbach`s common cause assumption*) (Valente, 2010; Pearl, 2000).

Ao considerar o processo amostral com base em um MEE com determinada estrutura causal, seria o mesmo que considerar resíduos independentes para cada variável resposta. Os resíduos representam o efeito dos pais da variável estudada que não estão no modelo. Assim, a covariância residual entre duas variáveis estudadas significa a existência de um pai comum entre elas. Por outo lado, sob suficiência causal, não há fonte de covariância residual, e uma estrutura diagonal é imposta à matriz de covariância (Valente, 2010).

Outra propriedade importante é a equivalência observacional de dois GADs diferentes, em que cada distribuição compatível com um GAD é também compatível com outro GAD equivalente. Se duas estruturas causais podem responder pela mesma descrição estatística multivariada, então a evidência estatística simplesmente não pode fazer distinção entre as duas estruturas. A equivalência estatística é o limite dos métodos de seleção de estruturas causais com base em dados observacionais. Estruturas equivalentes possuem as mesmas adjacências entre vértices e os mesmos *colliders*. Desta forma, $A \rightarrow C \rightarrow B$, $A \leftarrow C \leftarrow B$ e $A \leftarrow C \rightarrow B$ são equivalentes por induzirem o mesmo padrão de distribuição conjunta. Por outro lado, qualquer modificação em $A \rightarrow C \leftarrow B$ leva a uma estrutura não equivalente (Valente, 2010; Pearl, 2000).

5.2.7. Minimalidade e estabilidade

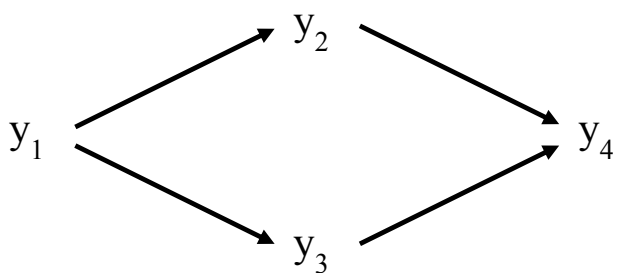
Em princípio, pode se ajustar a uma dada distribuição um grande número de estruturas causais. Para um conjunto de variáveis em suficiência causal, uma estrutura acíclica extremamente complexa pode, com uma parametrização finamente ajustada, mimetizar a distribuição gerada por várias outras estruturas. Torna-se necessário definir qual seria o critério para preferência de modelo. Geralmente, torna-se interessante desprezar teorias mais complexas quando existe uma teoria mais simples e igualmente consistente com as evidências (Pearl, 2000).

Assim, uma estrutura passa a ser preferida em relação a outra quando esta última, sob determinada parametrização, tem capacidade de mimetizar a distribuição gerada por um modelo sob a primeira estrutura, mas o contrário não é verdadeiro. A estrutura preferida tem menor poder expressivo ou flexibilidade, mas não necessariamente menos parâmetros, evitando o sobreajuste. O poder expressivo de uma estrutura é dado pelas independências nela presentes, sem se preocupar com a parametrização. Logo, é

interessante buscar por estruturas consistentes com a distribuição conjunta observada, que pode ser utilizada por modelos amostrais e gerar tal distribuição, com todas as suas dependências e que tenham o menor poder expressivo possível, ou seja, que sejam mínimas (Valente, 2010).

É possível que, dada uma estrutura causal, a distribuição conjunta de variáveis apresente independências condicionais além daquelas que seguem logicamente da Condição Causal de Markov (Valente, 2010).

Considere a estrutura a seguir, um exemplo utilizado por Spirtes et al. (2001):



Resultante desta estrutura causal, temos o seguinte MEE:

$$y_1 = e_1$$
$$y_2 = \lambda_{21} y_1 + e_2$$
$$y_3 = \lambda_{31} y_1 + e_3$$
$$y_4 = \lambda_{42} y_2 + \lambda_{43} y_3 + e_4.$$

Como consequência da Condição Causal de Markov, algumas independências condicionais estatísticas são esperadas, como por exemplo, y_2 e y_3 condicional a y_1, ou y_1 e y_4 condicional a y_3 e y_2. Todavia, é possível construir um modelo em que y_1 e y_4 sejam marginalmente independentes, o que não demonstra uma d-separação, uma vez que existem duas trilhas ativas entre os dois vértices no gráfico. Para tal, bastaria fazer com que $\lambda_{21}\lambda_{42} = -\lambda_{31}\lambda_{43}$, de modo que as duas trilhas se cancelem (Valente, 2010).

Na prática, esta situação raramente acontece e precisaria de um ajuste fino dos parâmetros de modo assintótico. Para não considerar esta situação, parte-se do pressuposto de que a distribuição conjunta é estável, ou crível (*faithful*) em relação ao modelo causal que a gera. A restrição de que todas as independências condicionais na distribuição conjunta resultante de um modelo causal são estáveis tendo como consequência a impossibilidade de se destruir independências estatísticas condicionais pela simples modificação dos valores dos parâmetros. No exemplo anterior, basta

modificar os valores do parâmetro de modo a tornar $\lambda_{21}\lambda_{42} = -\lambda_{31}\lambda_{43}$ falsa para destruir a independência marginal entre y_1 e y_4). Se uma independência condicional estatística ocorre para qualquer parametrização do modelo causal, esta independência é estável, como ocorre no caso das independências estatísticas que refletem d-separações em um modelo sob Condição Causal de Markov (Valente, 2010).

Pearl (2000) apresenta uma analogia interessante que permite observar a relação próxima entre minimalidade e estabilidade. Para tal, considera-se uma figura na qual é possível observar uma cadeira. É necessário decidir-se entre duas teorias:

T1 – O objeto na figura é uma cadeira

T2 – O objeto na figura pode ser uma cadeira ou pode ser duas cadeiras posicionadas de tal maneira que uma esconde a outra.

A preferência por T1 pode ser justificada baseando-se na minimalidade ou na estabilidade. Sob minimalidade, escolhe-se T1 porque o conjunto de cenas possíveis admitindo-se T1 é um subconjunto das cenas possíveis admitindo-se T2. Pelo menor poder expressivo de T1, escolhe-se esta teoria em detrimento de T2 para evitar um sobreajuste, a não ser que existam evidências favoráveis a T2. Por outro lado, a escolha de T1 pode-se basear na estabilidade, por considerar que o alinhamento perfeito entre dois objetos na figura é improvável e instável em relação a pequenas modificações no ângulo de visão.

5.2.8. Premissas

O algoritmo IC busca por estruturas causais para o modelo $\mathbf{y}_i = \mathbf{\Lambda}\mathbf{y}_i + \mathbf{e}_i$. De acordo com Spirtes et al. (2001), a seleção da classe de estruturas equivalentes é proposta sob as seguintes premissas (Valente, 2010):

1. O conjunto de variáveis estudadas é causalmente suficiente.
2. Cada unidade na população apresenta as mesmas relações causais entre variáveis.
3. A distribuição das variáveis observadas é crível em relação a um GAD.
4. As decisões estatísticas exigidas pelo algoritmo são corretas para a população.

A segunda premissa expressa o fato de que a conexão entre estruturas causais e distribuições pode se tornar problemática em situações em que se tem uma distribuição

conjunta composta de distribuições mistas, cada uma sendo relativa a uma subpopulação com uma estrutura causal específica (cada uma delas compatível com uma estrutura causal diferente). A quarta premissa indica que não há garantias de recuperação da classe correta das estruturas causais se as decisões estatísticas a respeito de independências condicionais, com base em uma amostra não correspondem ao que ocorre na população. Podem ocorrer erros, como aqueles provenientes de efeito amostral (Valente, 2010).

As premissas não são mais fortes do que as que geralmente são aplicadas na utilização de MEEs no contexto de modelos mistos em genética quantitativa. A primeira premissa é, provavelmente, a mais forte, Tendo como consequência prática imporá imposição de uma distribuição independente para os resíduos do MEE, ou seja, impor uma estrutura diagonal para a matriz de covariância residual. A imposição de tal estrutura é comum na aplicação recente de MEEs, sob a justificativa de atingir a identificabilidade dos parâmetros. Uma premissa forte adicional nas aplicações de MEEs é o conhecimento prévio a respeito da estrutura causal. Esta premissa pode ser evitada se utilizarmos o algoritmo IC para explorar o espaço de estruturas causais sem modificar a construção feita para os resíduos do modelo (Valente, 2010).

A quarta premissa é evidentemente necessária para garantir a recuperação de uma estrutura causal crível em relação à densidade conjunta populacional, mas erros estatísticos não resultam necessariamente em erros na busca. Tal situação ocorre quando existem mais de um conjunto de vértices que d-separam um par específico de vértices, mas nem todos refletem independência estatística na distribuição de probabilidade obtida de uma amostra (Spirtes et al., 2001).

5.2.9. Algoritmo IC

O algoritmo IC é utilizado para selecionar estruturas causais (ou uma classe de estruturas observacionalmente indistinguíveis) a partir das associações observadas entre as características. Este algoritmo se baseia em uma série de perguntas a respeito da independência condicional estatística entre as variáveis e na pressuposição de que estas independências são reflexos de d-separações na estrutura causal. Os dados de entrada do algoritmo são os elementos de uma matriz de correlação, dos quais podem se avaliar dependências condicionais. A saída do algoritmo é um gráfico parcialmente direcionado, sendo este um gráfico que contém linhas direcionadas e não direcionadas, que representa

uma classe de estruturas causais compatíveis com as independências condicionais obtidas. Esta classe funciona como uma restrição do espaço de hipóteses causais em relação ao espaço inicial. Assim, o algoritmo tenta construir um conjunto de GADs que satisfazem um dado conjunto dado de d-separações, isso, se tais GADs existem (Spirtes et al., 2001).

Considerando **V** um conjunto de variáveis aleatórias, o algoritmo IC consiste nos seguintes passos (Valente et al., 2011; Rosa et al., 2011; Valente, 2010):

1 – Para cada par de variáveis A e B em **V**, procure por um conjunto de variáveis $\mathbf{S}_{AB}$ de modo que A seja independente de B condicionalmente a $\mathbf{S}_{AB}$. Se A e B são dependentes condicionalmente a qualquer um dos possíveis grupos de variáveis remanescentes, conecte A e B com uma linha não-direcionada. Esta etapa do algoritmo tem como resultado o gráfico não direcionado **U**.

2 – Para cada par de variáveis não adjacentes A e B com uma variável adjacente em comum C em $\boldsymbol{U}$ (isto é, $A – C – B$), procure por um conjunto de variáveis $\mathbf{S}_{AB}$ que contém C de modo que A seja independente de B dado $\mathbf{S}_{AB}$. Se tal conjunto não existe, oriente as linhas da estrutura estudada em direção a C ($A \rightarrow C \leftarrow B$). Caso o conjunto exista, continue.

3 – No gráfico parcialmente direcionado resultante da etapa anterior, oriente ao máximo as linhas restantes, de maneira que não apareçam ciclos ou *colliders* além daqueles previamente identificados.

O objetivo do primeiro passo do algoritmo é obter um gráfico que especifique pares de características que são diretamente conectadas, mas sem especificar direção causal. Vértices adjacentes em um gráfico não são d-separados condicionalmente a qualquer conjunto de vértices remanescentes. A consequência observacional é que estas variáveis são estatisticamente dependentes condicionalmente a qualquer conjunto de variáveis restantes (Rosa et al., 2011; Valente et al., 2011; Valente, 2010).

O segundo passo tem como objetivo orientar linhas, por intermédio da busca por vértices no gráfico no qual setas convergem de ambos os sentidos em uma trilha (*colliders*). Estruturas internas dos gráficos compostas por um *collider* que sofre influência causal de dois vértices que não são conectados são chamadas *unshielded*

colliders (Spirtes et al., 2001), como em $A \rightarrow C \leftarrow B$. Nesta estrutura, os pais são d-separados condicionalmente a pelo menos um conjunto de variáveis restantes no gráfico completo, mas não se o vértice *C* pertence a este conjunto. Condicionalmente a C, a trilha entre A e B por intermédio de C permite o fluxo de dependência, não ocorrendo d-separação. A consequência observacional é que A e B nunca são estatisticamente independentes condicionalmente a qualquer conjunto de variáveis que contenha C.

Já o terceiro passo consiste em orientar linhas adicionais de maneira que isso não resulte em um novo *collider* ou em um ciclo. Desta forma, se três variáveis em um gráfico semi-direcionado hipotético resultante do passo 2 se apresentam conectados como em $A \rightarrow B - C$ a linha entre B e C deve ser orientada em direção a C, uma vez que a direção contrária resulta em um *collider* em B que não foi detectado no passo anterior. O mesmo deve ser feito para uma linha caso uma das direções resulte em um ciclo (Valente, 2010).

5.3. Interpretação dos parâmetros

Segundo Valente (2010), os parâmetros de locação (efeitos genéticos aditivos) e os parâmetros de dispersão (componentes de variância e covariância genética) são levados em consideração tanto no MEE quanto no MMC padrão. Contudo, a interpretação paramétrica se modifica de acordo com o modelo. Considere como exemplo, dados provenientes de um modelo amostral bicaracterística recursivo, no qual uma característica A tem efeito causal na característica B. Sob modelo recursivo, a correlação genética representa a associação linear entre dois efeitos genéticos aditivos não-observáveis, cada um afetando diretamente uma característica específica. Porém, caso um modelo multicaracterística seja escolhido para representar esse sistema multivariado, esta não seria a única fonte de variação, pois existe no modelo amostral uma associação indireta entre o efeito genético de A e o fenótipo de B, ou seja, o efeito genético de A tem influência causal sobre o fenótipo de A, que por sua vez, tem influência sobre B (é pai de B), na estrutura causal descrita. A correlação genética, em um modelo recursivo, não considera este efeito indireto. No entanto, a correlação genética em um modelo multicaracterística representa toda a associação de origem genética, independentemente de esta ser considerada direta ou indireta no contexto recursivo. Por este motivo, torna-se

concebível que a correlação genética sob modelo multicaracterísticas seja diferente de zero mesmo se os valores genéticos aditivos forem independentes no contexto recursivo.

5.4. Processo de estimação de um MEE

Kline (2011) descreve todas as etapas que devem ser percorridas por um pesquisador, para que este possa construir e testar um MEE. Primeiro, é necessário especificar as relações entre variáveis que compõem o modelo, onde as hipóteses da pesquisa são expressas na forma de um sistema de equações. É possível iniciar o processo de especificação com o desenho do diagrama, que posteriormente, é traduzido em uma série de equações, nas quais são definidos os parâmetros que serão estimados pelo modelo.

Na etapa seguinte é determinado se o modelo é identificável. Um modelo é identificável se for teoricamente possível derivar uma estimativa única para cada um dos parâmetros que devem ser calculados. Assim como na resolução de um sistema de equações lineares, deve haver uma compatibilidade entre o número de parâmetros desconhecidos (estimados pelo modelo) e o número de parâmetros conhecidos, (relacionados com a quantidade de variáveis observáveis que o modelo). Contudo, não é tão simples, pois diferentes tipos de MEE atendem a requerimentos específicos para serem identificados. Se um modelo houver falha em atendê-los, ele não estará identificado, e as tentativas de estimá-lo podem não ser bem sucedidas (Kline, 2011).

O terceiro passo começa com a coleta e preparação das variáveis que fazem parte do modelo. Uma vez que se tenham essas variáveis, é possível analisar os dados. Nesse passo são usados programas computacionais capazes de calcular os parâmetros desejados, com base nos dados empíricos. Tal estimação, geralmente, é feita por meio da máxima verossimilhança (Codes, 2005).

Após uma primeira estimação do modelo hipotético, avalia-se o ajuste, o que determina o quão adequadamente o modelo explica os dados. Para Kline (2011), é frequente, e até esperado, que os primeiros modelos elaborados não se ajustem bem aos dados. Quando isso acontece, deve se reespecificar o modelo. Tal reespecificação deve reiniciar toda a trajetória que fora descrita, de modo que se chegue novamente à etapa de avaliação do ajuste do modelo que foi revisado e construído com base nos mesmos dados, e assim sucessivamente.

Quando se obtém a um resultado estável e bem ajustado, tem início a discussão dos achados. Logo, os processos de construção e teste de um MEE são uma forma convencional de abordagem de proceder à modelagem (Codes, 2005).

5.5. Verificação do ajuste de um MEE

Ao se avaliar os resíduos, a situação ideal seria aquela onde não há diferença entre as matrizes de covariâncias. Em situações reais, um ajuste de modelo é considerado satisfatório quando os resíduos são os mais próximos possíveis de zero. Essa diferença entre as matrizes residuais é comum, isso significa que algumas variâncias e covariâncias das variáveis observadas não são exatamente preditas pelo modelo (Codes, 2005).

A partir da comparação das matrizes de covariância, foram criados coeficientes que verificam o ajuste global de um MEE. Existem diversos índices de ajuste, e dentre os mais utilizados estão: "Bentler-Bonnet Normed Fit Index" (NFI), Comparative Fit Index (CFI), Bentler-Bonnet Non-Normed Fit Index (NNFI) e Root Mean Square Error of Approximation (RMSEA) (Codes, 2005).

Como citado anteriormente, tais índices se caracterizam por serem medidas de ajuste global, isto é, se referem à adequação do modelo como um todo. Porém, diferenciais de ajuste não são captados nem reportados, se uma parte da modelagem mostre uma melhor adequação aos dados do que outra. Já os coeficientes refletem diferentes facetas quanto ao ajuste de um modelo. Desta forma, é recomendado que a adequação de um modelo aos dados seja aferida pelo conjunto de índices (Codes, 2005; Kline, 2011).

5.6. Aplicações dos MEEs sob o contexto do melhoramento animal

A utilização de MEEs no contexto de modelos mistos em genética quantitativa foi introduzida por Gianola e Sorensen (2004), que apresentaram uma metodologia para inferência de MEE e estudaram o relacionamento causal entre fenótipos e parâmetros genéticos. A partir de então, muitos autores utilizaram tais modelos para estudar sistemas multicaracterísticas.

De los Campos et al. (2006a; 2006b) utilizaram os MEEs na avaliação da produção de leite e contagem de células somáticas (CCS) em caprinos de leite e em bovinos de leite, respectivamente. Porém, não foram consideradas as informações de pedigree devido à limitação do programa LISREL. A metodologia empregada nestes estudos produziu resultados que indicam que a associação negativa entre as duas características tem como causa mais importante o efeito negativo da enfermidade na produção de leite e não um efeito de diluição das células somáticas no leite, quando a produção é maior (Rosa e Valente, 2012; Valente, 2010).

O relacionamento entre CCS e produção de leite também foi estudado por Wu et al. (2007), que utilizaram uma extensão do modelo apresentado por Gianola e Sorensen (2004). Eles propuseram um modelo que assumiu heterogeneidade de relações de recursividade e *feedback* entre diferentes níveis do banco de dados utilizado, o que também resulta em diferentes estimadores de componentes de variância para cada nível. A análise foi feita utilizando o programa SirBayes e os resultados obtidos indicaram efeitos de maior magnitude da CCS sobre a produção de leite, e efeitos de menor magnitude para o sentido inverso.

Para investigar o relacionamento entre tamanho de leitegada e peso médio dos leitões, em suínos Landrace e Yorkshire, Varona et al. (2007) utilizaram um modelo recursivo. Os autores propuseram uma estrutura causal na qual o peso médio dos leitões é influenciado pelo tamanho da leitegada. Houve ausência de efeitos recursivos em suínos Landrace e presença de recursividade apenas entre resíduos para suínos Yorkshire.

Wu et al. (2008) estudaram o relacionamento causal entre características lineares e de limiar utilizando uma outra extensão dos modelos apresentados por Gianola e Sorensen (2004) por meio de modelos hierárquicos bayesianos. Neste trabalho foi estudado o relacionamento entre incidência de mastite clínica e a produção de leite em diferentes períodos de gestação. Os resultados indicam que a incidência de mastite clínica tem um efeito negativo sobre a produção de leite, dentro de um mesmo período e a produção de leite em um período em que há ocorrência de mastite tem um efeito de baixa magnitude sobre subsequente. Dessa forma estes efeitos tendem a diminuir à medida que o animal está em períodos mais avançados de gestação (Rosa et al., 2011; Valente 2010).

Wu et al. (2008) propuseram um modelo de limiar-Gaussiano para inferir o relacionamento recursivo e simultâneo entre caracteres binários e Gaussianos, e utilizaram como uma metodologia com o mesmo sistema do MEE, para estudar os relacionamentos entre mastite clínica e produção de leite (Rosa et al., 2011).

O relacionamento causal entre o intervalo de gestação, dificuldade de parto e mortalidade perinatal foi estudado por de Maturana et al. (2007) que empregou um modelo recursivo, combinando uma análise de população heterogênea de Wu et al. (2007) e análise de características de limiar (Wu et al., 2008). Os resultados indicaram um intervalo de gestação ótimo (274 dias) com respeito à dificuldade de parto e mortalidade perinatal (Valente, 2010). E König et al.(2008) relataram uma aplicação referente a MEE de limiar bayesiano para o relacionamento entre problemas de casco e produção de leite em gado holandês (Rosa et al., 2011).

A utilização de um MEE, através de uma abordagem bayesiana e com aplicação do algoritmo IC para a distribuição conjunta de dados fenotípicos de codornas europeias, foi realizada por Valente et al. (2011) com o objetivo de aplicar a metodologia proposta por Valente et al. (2010) à um conjunto de dados reais.

6. VALIDAÇÃO CRUZADA

É uma abordagem que permite estimar o quanto um modelo, do qual foi possível um aprendizado a partir de dados observados, poderá ser bom ao ser aplicado em dados futuros desconhecidos até o momento. É uma ferramenta utilizada, principalmente, em situações onde o objetivo é a predição (Patterson and Thompson, 1971; Nephawe, 2004).

Originalmente, a validação cruzada foi empregada para avaliar a validade preditiva de equações de regressão linear utilizadas para prever o critério de desempenho em um conjunto de testes (Browne, 2000). Este autor afirma ainda que os coeficientes de correlações múltiplas dentro da amostra original utilizada para atribuir valores aos pesos da regressão davam uma impressão optimista da efetividade preditiva da equação de regressão quando aplicadas para observações futuras.

Ao se aplicar a validação cruzada é necessário promover a separação da amostra de dados em subconjuntos complementares. Determina-se a elaboração do **conjunto de treinamento**, que é o subconjunto no qual é realizado a análise e, do **conjunto de validação** (ou conjunto de teste), que é o subconjunto através do qual a análise será validada.

Em aplicações da validação cruzada no contexto do melhoramento animal, os indivíduos apresentam graus de variados de relações genéticas. Logo, obter um conjunto de dados de treinamento e validação independentes é raramente possível (Pérez-Cabal et

al., 2012). Assim, a forma com que são construídos os conjuntos de treinamento e validação influencia nos resultados da validação cruzada, sendo que o nível de parentesco entre os indivíduos dos conjuntos é um fator relevante (VanRaden et al., 2009).

MATERIAL E MÉTODOS

Os dados utilizados na análise consistem de um banco de dados contendo informações completas de 1.286 fêmeas de codornas de corte (*Coturnix*), pertencentes a 10 gerações distintas, provenientes da Granja de Melhoramento de Aves, do Departamento de Zootecnia, da Universidade Federal de Viçosa, MG.

As características avaliadas foram: o peso do animal ao nascimento (PN), o peso do animal aos 35 dias de idade (P35), a idade ao primeiro ovo (IPO), o peso médio dos ovos dos 42 aos 182 dias de idade (PMO) e o número de ovos produzidos dos 42 aos 182 dias de idade (NO).

Baseando-se em conhecimentos prévios e na informação temporal, considerou-se que o modelo estrutural hipotético apresentado a seguir é o modelo que melhor explica o relacionamento causal entre PN, P35, IPO, PMO e NO.

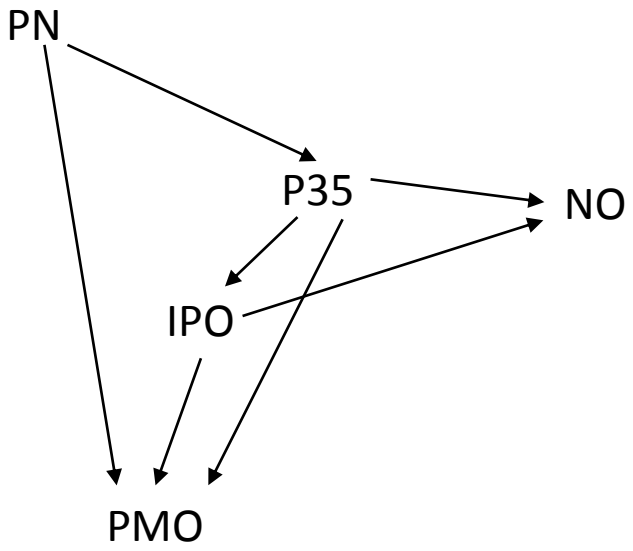


FIGURA 2. Modelo hipotético que considera o relacionamento recursivo entre as características PN, P35, IPO, PMO e NO.

São assumidas para as características as seguintes distribuições:

PN ~ N (μ_{PN},$\sigma^2_{e_{PN}}$);
P35 ~ N (μ_{P35},$\sigma^2_{e_{P35}}$);
IPO ~ N (μ_{IPO},$\sigma^2_{e_{IPO}}$);
PMO ~ N (μ_{PMO},$\sigma^2_{e_{PMO}}$);
NO ~ N (μ_{NO},$\sigma^2_{e_{NO}}$); e
NO ~ Poi (m).

O modelo em questão define que PN influencia de forma recursiva P35 e PMO, que P35 influencia da mesma maneira IPO, PMO e NO, e por sua vez IPO influencia NO.

As equações do modelo estrutural recursivo foram estimadas seguindo formulação de Modelos Lineares Generalizados Mistos. Tal modelo é especificado através de um preditor linear, uma função de ligação e por uma distribuição da família exponencial e ainda, permite a especificação de coeficientes aleatórios.

O preditor linear das equações do modelo para a variável y_i é dado por:

$$\eta = \mathbf{X\beta} + \mathbf{Zu} + \sum_{i=1}^{P} \lambda_{i,pi} \mathbf{y}_{pi} , \qquad [1]$$

onde:

η é o preditor linear;

X é a matriz do modelo que relaciona os efeitos fixos a η;

β é o vetor de efeitos fixos de geração;

Z é uma matriz do modelo que relaciona η a **u**;

u é o vetor de efeitos genéticos aditivos;

λ é o coeficiente estrutural;

$\mathbf{y}_{pi}$ são os pais da variável resposta; e

P é o número de pais da variável resposta.

A esperança condicional da resposta y dado **x**, **u** e $\mathbf{y}_{pi}$ é "ligada" ao preditor linear η através de uma função de ligação $g(.)$, que é definida por:

$$\eta = g\left(\mathrm{E}\left[\mathrm{y} \mid \mathbf{x},\ \mathbf{u}, \mathbf{y}_{pi}\right]\right). \qquad [2]$$

Ao se escolher uma distribuição da família exponencial para a distribuição condicional da variável resposta dada as variáveis explanatórias, os efeitos genéticos e os pais da variável aleatória, a especificação do modelo está completada. No panorama deste estudo, é proposto a abordagem das variáveis que são normalmente distribuídas e as variáveis relacionadas com dados de contagem. Estas abordagens são apresentadas a seguir:

a) **Variáveis Gaussianas**: para tais variáveis são tipicamente assumidas uma função de ligação identidade e uma distribuição normal. Onde, nesse caso,

$$y = \eta + e,$$

com e ~ $N(0,\sigma^2)$. Com densidade condicional

$$f\left(\mathrm{y} \mid \mathbf{x}, \mathbf{u}, \mathbf{y}_{pi}\right) = \sigma^{-1}\phi\left(\eta\sigma^{-1}\right), \qquad [3]$$

onde ϕ denota a densidade normal padrão.

b) **Contagens**: o modelo para dados de contagem, que deve ser inteira e positiva, é especificada através de uma função de ligação log e uma distribuição Poisson, correspondentemente:

$$\eta = \ln \mathrm{E}\left[\mathrm{y}_i \mid \mathbf{x}, \mathbf{u}, \mathbf{y}_{pi}\right]$$

e

$$f\left(\mathrm{y} \mid \mathbf{x}, \mathbf{u}, \mathbf{y}_{pi}\right) = \frac{\left[\exp\left(\eta\right)\right]^{y}}{y!}\exp\left(-\exp\left(\eta\right)\right). \qquad [4]$$

É definido que o modelo hipotético proposto segue todas as premissas do modelo de equações estruturais. Assume-se também os coeficientes estruturais iguais a 1, determinando igual magnitude dos efeitos das variáveis pais sobre as variáveis resposta. E assume-se ainda geração como o efeito fixo.

A comparação da equação onde NO tem distribuição normal com a equação onde NO assume distribuição Poisson foi realizada por meio dos critérios AIC e BIC.

Para o cálculo dos coeficientes da matriz de parentesco entre os animais, foi utilizado o procedimento INBREED, do software estatístico SAS versão 9.4 para Windows. Para as análises das equações do modelo recursivo foram realizadas utilizando o procedimento GLIMMIX do referido software (Apêndice I).

A validação dos resultados do modelo foi testada por meio da técnica de validação cruzada, a fim de determinar a Acurácia na Predição dos valores fenotípicos (AP). Dessa

forma, estabeleceu-se que o conjunto dos dados de treinamento corresponde ao conjunto inicial dos dados (1.286 fêmeas de codornas de corte) e que o conjunto de dados de validação é um conjunto de dados menor, com 1.257 fêmeas de codornas de corte, que contempla uma modificação dos dados originais, no qual 29 animais que constituíam a última geração tiveram fenótipo definidos como “desconhecido”. A AP corresponde à correlação entre os valores fenotípicos observados do conjunto de dados iniciais e os valores fenotípicos estimados a partir do conjunto de dados de validação, com valores entre 0 e 1.

RESULTADOS E DISCUSSÃO

Na Figura 3 estão representados os efeitos genéticos diretos u_{PN}, u_{P35}, u_{IPO}, u_{PMO}, e u_{NO}, onde cada um destes afetam diretamente PN, P35, IPO, PMO e NO, respectivamente. Pela metodologia de aplicação, não são consideradas as correlações genéticas entre os efeitos genéticos. Note que podem existir os efeitos genéticos indiretos que são mediados pelos fenótipos de outras características, como por exemplo, no caso de u_{PN}, que afeta P35 de forma indireta através do PN. Este fato é de extrema importância, pois os efeitos genéticos por si só não são suficientes para a determinação dos fenótipos, dessa forma esses efeitos devem ser associados com toda a informação que a estrutura causal pode oferecer em relação às características.

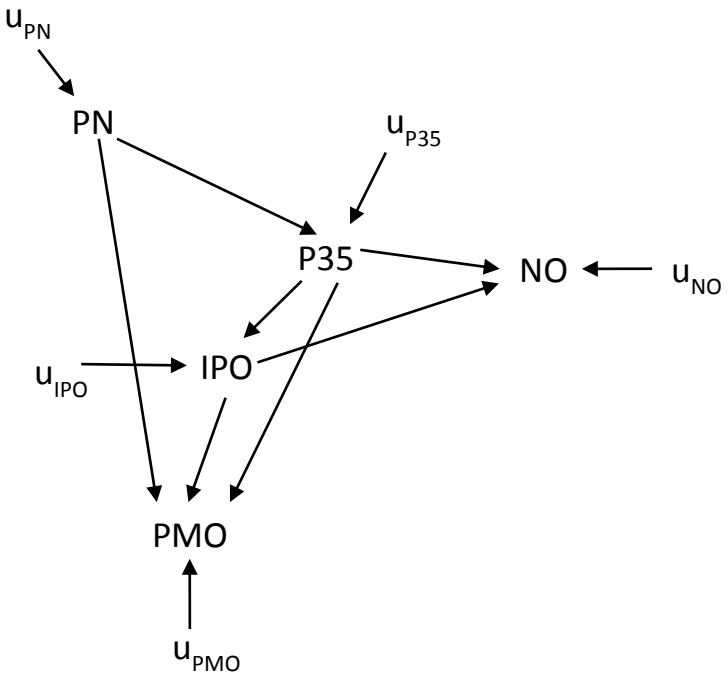


FIGURA 3. Modelo hipotético que considera o relacionamento recursivo entre as características PN, P35, IPO, PMO e NO, com representação dos efeitos genéticos aditivos diretos u_{PN}, u_{P35}, u_{IPO}, u_{PMO}, e u_{NO}.

Ao realizar, por exemplo, uma intervenção externa em P35, define-se seu valor como uma constante (c), como pode ser observado na Figura 4, isto irá provocar uma mudança direta em IPO, PMO e NO e acarretará em um efeito indireto sobre NO e PMO através do efeito de IPO. No entanto, nenhum efeito é aplicado sobre PN e nem sobre o valor genético de nenhuma das variáveis, uma vez que P35 não possui efeito sobre

nenhuma delas, Uma intervenção é definida por Pearl (2000) como uma mudança na relação funcional entre as características, forçando algumas variáveis a assumirem um valor fixo, que reflete numa "poda" do modelo e a substituição da variável manipulada por uma constante, como é representado na Figura 4, que mostra que as trilhas que ligam PN e u_{P35} a P35 são eliminadas.

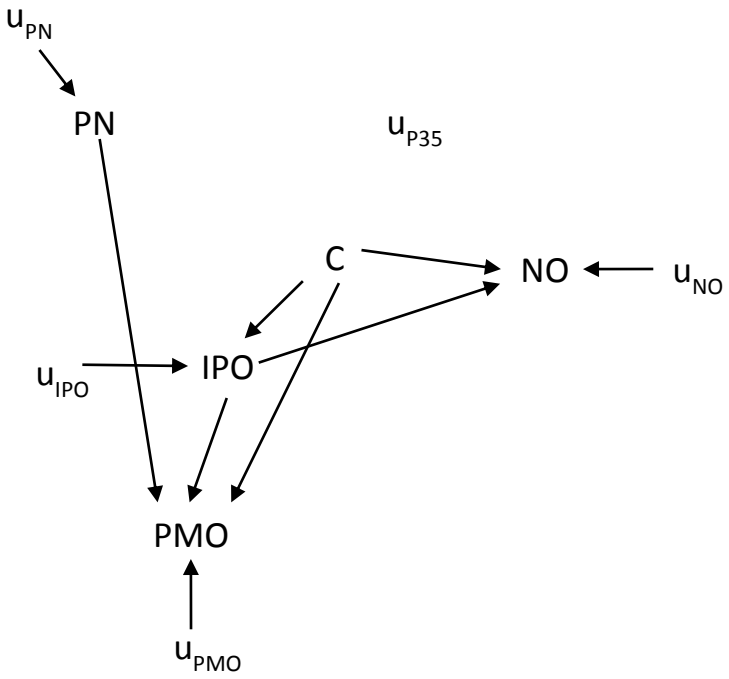


FIGURA 4. Modelo hipotético que considera o relacionamento recursivo entre as características PN, P35, IPO, PMO e NO, com representação dos efeitos genéticos aditivos diretos u_{PN}, u_{P35}, u_{IPO}, u_{PMO}, e u_{NO} e representação de uma possível intervenção externa (c) em P35.

Ainda como resultado da possível intervenção apresentada na figura anterior, determina-se desta maneira, que o efeito genético u_{P35} passa a não exercer efeitos indiretos sobre IPO, PMO e NO. Resultado que pode confirmar a afirmação de Valente et al. (2013) que diz que através da utilização da metodologia de equações estruturais pode ser possível predizer o mérito genético de indivíduos em um conjunto de tratamentos com magnitudes diferentes, em casos onde não existam um tratamento preferencial ou ainda, em casos em de diferentes níveis de tratamentos preferenciais, assumindo que não existe nenhuma fonte de associação causal entre as características.

Para a análise de validação cruzada, os dados do conjunto de validação foram analisados pelo modelo descrito com o objetivo de predizer fenótipos dos animais que tiveram o seu fenótipo como "desconhecido". Como resultado foi predito um valor

fenotípico para cada animal da última geração e estes foram comparados, por meio do coeficiente de correlação (AP), com o valor do fenótipo observado dos dados originais. Os coeficientes de correlação (Tabela 5) entre os fenótipos observados e os fenótipos preditos foram calculados e utilizado como uma medida de acurácia das predições dos valores fenotípicos.

Tabela 5. Correlações dos valores estimados com os valores observados, Acurácia na predição dos valores fenotípicos e a correlação entre estes dois critérios, para as características avaliadas.

Característica	Cor	AP
PN	0,96	0,96
P35	0,99	0,99
IPO	0,87	0,87
PMO	0,92	0,87
NO ~ N	0,45	0,35
NO ~ Poi	0,97	0,96

PN- peso ao nascimento; P35- peso aos 35 dias de idade, IPO- idade ao primeiro ovo; PMO- peso médio dos ovos; NO- número de ovos; N- distribuição normal; Poi- distribuição Poisson; Cor- correlação entre os valores estimados a partir do conjunto de dados de treinamento com os valores fenotípicos observados; AP- acurácia na predição dos valores fenotípicos.

É possível notar que os valores da acurácia de predição são considerados altos para todas as características com exceção do número de ovos que apresentou um valor (0,355) bem abaixo dos valores encontrados para todas as outras características. Isto sugere que a adoção de tal característica para o modelo recursivo apresentado não é interessante. Entretanto, vale a pena destacar a superioridade da utilização da distribuição de Poisson na estimação dos valores fenotípicos de NO.

Para Devijver e Josef (1982) e Pérez-Cabal et al. (2012), em uma validação cruzada eficiente, o tamanho do conjunto de dados é extremamente importante e, dependendo da característica, seriam necessários pelo menos 1.000 animais para se ter certeza de alcançar acurácias comparáveis. De acordo com o exposto por tais autores, pode se afirmar, que a validação cruzada se apresenta-se como uma ferramenta eficiente na representação da acurácia de predição.

Geralmente os valores de acurácia na predição variam amplamente. Os valores de acurácia mais baixos estão, geralmente associados às características de baixa

herdabilidade, como no caso de NO. Se o valor de acurácia de uma característica for próximo ou superior a 0,75, é um bom indicativo de que uma possível seleção de indivíduos com base em tal característica seria bem aplicada (VanRaden, 2008). Os valores de acurácia na predição dos valores fenotípicos encontrados neste estudo indicam que o método proposto pode ser considerado eficiente no ajuste do modelo proposto.

A Tabela 6 mostra que a correlação dos valores estimados com os valores observados de NO é considerada alta (0,97) para o modelo considerando NO ~ Poi. Da mesma forma, os critérios AIC e BIC foram menores considerando esta distribuição, com valores de 1030,37 e 1035,53, respectivamente, para o conjunto de dados de treinamento e 1011,97 e 1017,11, respectivamente, para o conjunto de dados de validação.

Tabela 6. Critérios de comparação de NO nas distribuições normal e Poisson

Critérios de comparação	Distribuições	
	NO ~ N	NO ~ Poi
Conjunto de dados de treinamento		
Cor	0,458	0,9715
AIC	11738,70	1030,37
BIC	11749,02	1035,53
Conjunto de dados de Validação		
AP	0,355	0,9721
AIC	11427,60	1011,97
BIC	11437,87	1017,11

NO- número de ovos; N- distribuição normal; Poi- distribuição Poisson; Cor- correlação dos valores estimados a partir dos dados do conjunto de treinamento com os valores observados de NO; AP- acurácia na predição dos valores fenotípicos em relação a NO; AIC- Akaike Information Criterion; BIC- Bayesian Information Criterion.

Diante do exposto, o modelo que considera NO com distribuição Poisson pode ser considerado o mais indicado do que o modelo que considera NO com distribuição normal.

Foram estimadas as variâncias genéticas aditivas e as estimativas de herdabilidade do modelo recursivo para os conjuntos de dados de treinamento (Tabela 7) e para o conjunto de dados de validação (Tabela 8).

Tabela 7. Estimativas de variância genética aditiva (σ_a^2) e herdabilidade (h^2) para as características do modelo para os dados de treinamento.

Características	σ_a^2	h^2
PN	0,4764	0,68
P35	643,48	0,96
IPO	11,9237	0,28
PMO	199,71	0,51
NO ~ N	32,669	0,05
NO ~ Poi	0,1655	0,0003

PN- peso ao nascimento; P35- peso aos 35 dias de idade, IPO- idade ao primeiro ovo; PMO- peso médio dos ovos; NO- número de ovos; N- distribuição normal; Poi- distribuição Poisson.

Ao se avaliar estes parâmetros, pode-se observar que as estimativas de herdabilidade para PN foram de 0,68 para ambos conjuntos de dados, sendo superiores ao valor relatado por Silva et al. (2013), que realizando análise unicaracterística para a estimação dos componentes de variância, que foi de 0,53. Estes valores também foram maiores do que o valor encontrado por Teixeira et al. (2013), que também estimaram os componentes de variância por meio de análises unicaracterísticas, utilizando o método da máxima verossimilhança restrita, que foi de 0,64, para uma das linhagens consideradas em seu estudo. Entretanto, estes autores também encontraram o mesmo valor de herdabilidade (0,68) para uma segunda linhagem.

Tabela 8. Estimativas de variância genética aditiva (σ_a^2) e herdabilidade (h^2) para as características do modelo para os dados de validação.

Características	σ_a^2	h^2
PN	0,4811	0,68
P35	518,78	0,89
IPO	12,2957	0,29
PMO	0,3957	0,42
NO ~ N	0,2769	0,001
NO ~ Poi	0,1655	0,0003

PN- peso ao nascimento; P35- peso aos 35 dias de idade, IPO- idade ao primeiro ovo; PMO- peso médio dos ovos; NO- número de ovos; N- distribuição normal; Poi- distribuição Poisson.

O peso aos 35 dias de idade apresentou valores de herdabilidade de 0,96 e 0,89, sendo superiores a 0,29 encontrado por Silva et al., (2013) e, 0,35 e 0,44 por Teixeira et al. (2013). A herdabilidade para IPO foram de 0,28 e 0,29, também superiores ao valor encontrado por Teixeira et al. (2012) que foi de 0,10. Os valores das herdabilidade para NO, 0,05 0,001 são considerados muito baixos e o valor encontrado a partir dos dados do grupo de treinamento se assemelham aos valores de herdabilidade para a mesma característica do trabalho de (Silva et al., 2013) que encontraram valores de herdabilidade entre 0,05 e 0,04 para duas linhagens diferentes de codornas de corte, e dos valores (0,03 e 0,01) apresentados por Teixeira et al. (2013) para taxa de postura.

Trabalhos futuros que permitam a comparação e a escolha de modelos que apliquem a metodologia proposta neste trabalho, que melhor representem o relacionamento causal entre as características de interesse são sugeridos.

CONCLUSÕES

Ao se considerar o relacionamento causal entre as características é possível entender a natureza e a ação dos efeitos genéticos aditivos sobre as variáveis estudadas.

O modelo de equações estruturais lineares generalizados mistos é eficiente na descrição fenotípica, visto que apresenta valores altos para a acurácia na predição de valores fenotípicos.

A avaliação do número de ovos, considerando a distribuição Poisson, foi melhor do que considerando tal característica normalmente distribuída.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. Information theory and an extension of the maximum likelihood principle. In: PETROV, B. N. e CSAKI, F., Proceedings of 2nd International Symposium on Information Theory, 1973. Budapest. Akademiai Kiado. p.267-281.

BRESLOW, N. E.; CLAYTON, D. G. Approximate Inference in Generalized Linear Mixed Models. **Journal of the American Statistical Association,** v. 88, n. 421, p. 9-25, 1993/03/01 1993.

BROWNE, M. W. Cross-Validation Methods. **Journal of Mathematical Psychology,** v. 44, n. 1, p. 108-132, 3// 2000.

CODES, A. L. M. MODELAGEM DE EQUAÇÕES ESTRUTURAIS: um método para a análise de fenômenos complexos. **Caderno CRH,** v. 18, n. 45, p. 484, 2005.

COSTA, T. R. **Modelos Lineares Mistos: Uma aplicação na produção de leite de vacas da raça Sindi**. 2010. 64 (Mestrado). Biometria e Estatística Aplicada, Universidade Federal de Pernambuco, Recife, PE.

DAVIS, C. S. **Statistical methods for the analysis of repeated measurements**. Springer, 2002. ISBN 0387953701.

DE LOS CAMPOS, G. et al. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. **J Anim Sci,** v. 84, n. 11, p. 2934-41, Nov 2006.

DE LOS CAMPOS, G.; GIANOLA, D.; HERINGSTAD, B. A structural equation model for describing relationships between somatic cell score and milk yield in first-lactation dairy cows. **J Dairy Sci,** v. 89, n. 11, p. 4445-55, Nov 2006.

DE MATURANA, E. L. et al. Analysis of fertility and dystocia in Holsteins using recursive models to handle censored and categorical data. **J Dairy Sci,** v. 90, n. 4, p. 2012-24, Apr 2007.

DEMÉTRIO, C. G. B. **Modelos Lineares Generalizados em Experimentação Agronômica**. 2002.

DEVIJVER, P. A.; KITTLER, J. **Pattern recognition : a statistical approach**. Englewood Cliffs [etc.]: PHI, 1982. ISBN 0136542360.

GIANOLA, D.; FOULLEY, J. Sire evaluation for ordered categorical data with a threshold model. **Genetics Selection Evolution,** v. 15, n. 2, p. 201 - 224, 1983.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. **Genetics,** v. 167, n. 3, p. 1407-24, Jul 2004.

HAASE, R. F. **Multivariate General Linear Models**. Sager, 2011. ISBN 9781412972499.

HARVILLE, D. Extension of the Gauss-Markov theorem to include the estimation of random effects. **The Annals of Statistics**, p. 384-395, 1976.

JÖRESKOG, K. G. A general method for estimating as linear structural equation system. In: GOLDBERGER, A. S. D., OTIS DUDLEY (Ed.). **Structural equation models in the social sciences**. New York, U.A.: Seminar Press, v.1970, 1973. p.85-112.

KEESLING, J. W. **Maximum likelihood approaches to causal flow analysis**: University of Chicago 1972.

KLINE, R. **Principles and practice of structural equation modeling**. Guilford Press, 2011. ISBN 9781606238776.

LITTELL, R. et al. **SAS for mixed models**. SAS Press, 2006. 840 ISBN 9781590475003.

MARTINS, E. N. et al. **Modelo Linear Misto**. Cadernos Didáticos. Viçosa, MG: Editira UFV. 38**:** 46 p. 1998.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. Chapman & Hall/CRC, 1989. ISBN 0412317605.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General),** v. 135, n. 3, p. 370-384, // 1972.

NEPHAWE, K. Application of random regression models to the genetic evaluation of cow weight in Bonsmara cattle of South Africa. **South African Journal of Animal Science,** v. 34, n. 3, p. p. 166-173, 2004.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika,** v. 58, n. 3, p. 545-554, 1971.

PEARL, J. Causal diagrams for empirical research. **Biometrika,** v. 82, n. 4, p. 669-688, 1995.

PEARL, J. **Causality: models, reasoning and inference**. Cambridge Univ Press, 2000.

PEARL, J. Causal inference in statistics: An overview. **Statistics Surveys,** v. 3, p. 96-146, 2009.

PÉREZ-CABAL, M. A. et al. Accuracy of genome enabled prediction in a dairy cattle population using different cross-validation layouts. **Frontiers in Genetics,** v. 3, 2012-February-28 2012.

PINHEIRO, J. C.; BATES, D. M. **Mixed-effects models in S and S-PLUS**. Springer, 2000. ISBN 0387989579.

RIOS-NETO, E. L.; OLIVEIRA, A. Aplicação de um modelo de idade-período-coorte para a atividade econômica no Brasil metropolitano. **Pesquisa e Planejamento Econômico,** v. 29, n. 2, p. 243-272, 1999.

ROCHA, E. B. D. et al. Aplicação dos modelos lineares generalizados na análise do número de estômatos em coentro (Coriandrum sativum L.): estimação bayesiana utilizando INLA (pp. 212-216). **Revista da Estatística da Universidade Federal de Ouro Preto,** v. 3, n. 3, 2014.

ROSA, G. J.; VALENTE, B. D. Inferring causal effects from observational data in livestock. **J Anim Sci**, Dec 10 2012.

ROSA, G. J. et al. Inferring causal phenotype networks using structural equation models. **Genet Sel Evol,** v. 43, p. 6, 2011.

ROSENBAUM, P. R.; RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. **Biometrika,** v. 70, n. 1, p. 41-55, 1983.

SCHUMAKER, R. E.; LOMAX, R. G. **A Beginer's Guide to Structural Equation Modeling**. 2. Mahwah, New Jersey - London Lawrence Erlbaum Associates, 2004. 498

SCHWARZ, G. Estimating the Dimension of a Model. **The Annals of Statistics,** v. 6, n. 2, p. 461-464, 1978.

SEARLE, S. R. **Linear models**. John Wiley & Sons, 1997. ISBN 1118491777.

SHIPLEY, B. **Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference**. Cambridge University Press, 2002. ISBN 0521529212.

SILVA, L. P. et al. Genetic parameters of body weight and egg traits in meat-type quail. **Livestock Science,** v. 153, n. 1–3, p. 27-32, 5// 2013.

SPEARMAN, C. The Proof and Measurement of Association between Two Things. **The American Journal of Psychology,** v. 15, n. 1, p. 72-101, 1904.

SPEARMAN, C. **The abilities of man; their nature and measurement**. New York: Macmillan Co., 1927.

SPIRTES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, Prediction, and Search**. The MIT Press, 2001. ISBN 0262194406.

TEIXEIRA, B. B. et al. Herdabilidade de características de produção e postura em matrizes de codornas de corte. **Ciência Rural,** v. 43, n. 2, p. 361-365, 2013.

TEIXEIRA, B. B. et al. Estimação dos componentes de variância para as características de produção e de qualidade de ovos em matrizes de codorna de corte. **Ciência Rural,** v. 42, n. 4, p. 713-717, 2012.

TEMPELMAN, R. J. Generalized linear mixed models in dairy cattle breeding. **Journal of dairy science,** v. 81, n. 5, p. 1428-1444, 1998.

THOMPSON, R. Sire evaluation. **Biometrics**, p. 339-353, 1979.

THURSTONE, L. L. Current issues in factor analysis. **Psychological Bulletin,** v. 37, n. 4, p. 189, 1940.

TUFTE, E. R. **The cognitive style of PowerPoint**. Graphics Press Cheshire, CT, 2003.

TURKMAN, M. A. A.; SILVA, G. L. **Modelos Lineares Generalizados - da teoria à prática**. 2003. 151

VALENTE, B. D. **Busca por estruturas causais no contexto de modelos mistos em genética quantitativa**. 2010. 63 (Doctor Science). Departamento de Ciência Animal, Universidade Federal de Minas Gerais, Escola de Vetérinaria - Universidade Federal de Minas Gerais.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. **Genetics,** v. 185, n. 2, p. 633-44, Jun 2010.

VALENTE, B. D. et al. Is structural equation modeling advantageous for the genetic improvement of multiple traits? **Genetics,** v. 194, n. 3, p. 561-72, Jul 2013.

VALENTE, B. D. et al. Searching for phenotypic causal networks involving complex traits: an application to European quail. **Genet Sel Evol,** v. 43, p. 37, 2011.

VANRADEN, P. Efficient methods to compute genomic predictions. **Journal of dairy science,** v. 91, n. 11, p. 4414-4423, 2008.

VANRADEN, P. M. et al. Invited review: reliability of genomic predictions for North American Holstein bulls. **J Dairy Sci,** v. 92, n. 1, p. 16-24, Jan 2009.

VARONA, L.; SORENSEN, D.; THOMPSON, R. Analysis of litter size and average litter weight in pigs using a recursive model. **Genetics,** v. 177, n. 3, p. 1791-9, Nov 2007.

VAZQUEZ, A. I. et al. Technical note: an R package for fitting generalized linear mixed models in animal breeding. **J Anim Sci,** v. 88, n. 2, p. 497-504, Feb 2010.

VENEZUELA, M. K. **Modelos lineares generalizados para análise de dados com medidas repetidas**. 2003. Universidade de São Paulo

WILEY, D. E. **The Identification Problem for Structural Equation Models with Unmeasured Variables**. Structural Models in the Social Sciences, A.S. Goldberger and O.D: Academic Press 1973.

WOLFINGER, R.; O'CONNELL, M. Generalized linear mixed models a pseudo-likelihood approach. **Journal of Statistical Computation and Simulation,** v. 48, n. 3-4, p. 233-243, 1993/12/01 1993.

WRIGHT, S. On the Nature of Size Factors. **Genetics,** v. 3, n. 4, p. 367-74, Jul 1918.

WRIGHT, S. Correlation and Causation. **Journal of Agricultural Research,** v. 20, p. 557-587, 1921.

WRIGHT, S. The Method of Path Coefficients. **Annals of Mathematical Statistics,** v. 5, p. 161-215, 1934.

WU, X. L. et al. Inferring relationships between somatic cell score and milk yield using simultaneous and recursive models. **J Dairy Sci,** v. 90, n. 7, p. 3508-21, Jul 2007.

WU, X. L.; HERINGSTAD, B.; GIANOLA, D. Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model. **Genet Sel Evol,** v. 40, n. 4, p. 333-57, Jul-Aug 2008.

Apêndice I – Rotina em SAS para a execução de MEEGM

```
/*Input dos dados*/

data codorna;
input animal 1-8 pai 9-16 mae 17-24 ger PN  P35  IPO  PMO    NO;
datalines;
6340030  503764  503894   1   8.7   250.6    67   12.92     115
6340032  503856  503462   1   9.3   254.2    64   12.91     110
6340033  503739  503748   1   8.3   196.5    69   12.42     112
6340042  503859  503693   1   7.5   249.2    61   14.22     101
6340052  503420  503810   1   9.7   203.5    71   10.5      106
6340081  503662  503658   1   7.3   192.7    64   12.44     113
6340085  503859  503693   1   7.5   208.8    75   11.47      94
6340115  503772  503738   1   8.1   209.9    70   13.02     111
6340140  503873  503704   1   9.6   220.7    64   13.32     116
6340142  503500  503486   1  10.5   243.4    68   13.93     109

 .
 .
 .

92111340 91105083 91107057 10 10.58  312.48   47   15.73     134
92111361 91105415 91106973 10 10.09  302.84   51   14.57     100
92111362 91106421 91105808 10 10.81  300.36   42   14.66     131
92111402 91106813 91106507 10 10.33  291.72   51   17.08     123
92111403 91105731 91106824 10  8.3   291.44   49   12.87     109
92111419 91106431 91106263 10  9.04  292.3    52   13.21     118
92111422 91106670 91105536 10 10.59  303.75   52   13.97     128
92111423 91105712 91105686 10 10.38  301.1    52   14.99      71
92111429 91106813 91106485 10  8.91  290.85   49   15.7      125
92111436 91105409 91106887 10   9.5  271.59   51   14        131
92111443 91106945 91105350 10 10.15  293.29   50   14.39      75
;

data validacao;

6340030  503764  503894   1   8.7   250.6    67   12.92     115
6340032  503856  503462   1   9.3   254.2    64   12.91     110
6340033  503739  503748   1   8.3   196.5    69   12.42     112
6340042  503859  503693   1   7.5   249.2    61   14.22     101
6340052  503420  503810   1   9.7   203.5    71   10.5      106
6340081  503662  503658   1   7.3   192.7    64   12.44     113
6340085  503859  503693   1   7.5   208.8    75   11.47      94
6340115  503772  503738   1   8.1   209.9    70   13.02     111
6340140  503873  503704   1   9.6   220.7    64   13.32     116
6340142  503500  503486   1  10.5   243.4    68   13.93     109

.
.
.

92111361 91105415 91106973 10    .     .     .     .     .
92111362 91106421 91105808 10    .     .     .     .     .
92111402 91106813 91106507 10    .     .     .     .     .
```

```
92111403 91105731 91106824 10    .     .     .     .     .
92111419 91106431 91106263 10    .     .     .     .     .
92111422 91106670 91105536 10    .     .     .     .     .
92111423 91105712 91105686 10    .     .     .     .     .
92111429 91106813 91106485 10    .     .     .     .     .
92111436 91105409 91106887 10    .     .     .     .     .
92111443 91106945 91105350 10    .     .     .     .     .
;
run;

/*Edição do Pedigree*/

data ped;
set codorna;
if pai eq 0 then pai=.;
if mae eq 0 then mae=.;
keep animal pai mae;
run;

data pai;
set codorna;
animal=pai;
pai=.;
mae=.;
keep animal pai mae;
run;

data mae;
set codorna;
animal=mae;
pai=.;
mae=.;
keep animal pai mae;
run;

/* Cálculo dos coeficientes da Matriz de parentesco */

data pedigree;
set ped pai mae;
sire=pai;
dam=mae;
run;

proc sort data=pedigree;
by animal;
run;

data pedigree;
set pedigree;
if animal ne lag(animal);
run;

proc inbreed data=ped covar outcov=amatrix;
var animal pai mae;
run;
```

```
/* Montagem da Matriz A para o LDATA */
data L2DATA;
set amatrix;
parm = 1;
row = _n_;
run;


/* Edição da Matriz de Parentesco, mantendo na matriz somente os
animais que tem informação de fenótipos */

data f0; set l2data; keep animal; run;

proc sort data=codorna; by animal;run;
proc sort data=l2data; by animal;run;

data f; merge codorna (in=s1) l2data (in=s2); if s1 & s2; by
animal; run;

data f1; set f; keep COL1-COL2014;run;
data f11; set f; keep animal ;run;

proc transpose data=f1 out=f2;
run;

data f3; merge f0 f2; drop _NAME_; run;

proc sort data=f3;by animal; run;
proc sort data=f11;by animal; run;

data f4; merge f11 (in=s1) f3 (in=s2); if s1 & s2;by animal;
run;

data f5; set f4; drop animal;
parm = 1;
row = _n_;
run;


/* Ajuste dos modelos e estimação dos parâmetros */

/* O código abaixo aplicado para a característica PN, para
ajuste do modelo também foram aplicados para as características
P35, IPO, PMO e NO (com distribuição normal.*/

Title 'Analise PN';
proc glimmix data=codorna noprofile IC=Q;
class animal ger;
model pn = ger /dist=normal link=identity;
random animal/type=lin(1) LDATA=f5 solution;
parms (0.4764) (0.2189);
ods output solutionr = solutionr;
output out=pnpredito pred=pnpredito;
run;
```

```
/* O código abaixo aplicado para a característica NO com
distribuição Poisson. */


Title 'Analise NO - poisson';
proc glimmix data=codorna noprofile IC=Q;
class animal ger;
model no = ger /dist=poisson link=log;
random animal p35 ipo /type=lin(1) LDATA=f5 solution ;
*output out=saidares resid=r pred=predicted;
parms (35)/;
ods output solutionr = blupnopoi;
output out=nopoipredito pred=nopoipredito;
run;


/* Validação cruzada */

/* O código abaixo aplicado para a característica PN, para
ajuste do modelo, para os dados do conjunto de validação, também
foram aplicados para as características P35, IPO, PMO e NO (com
distribuição normal.*/


Title 'Analise PN-cross';
proc glimmix data=validacao noprofile IC=Q;
class animal ger;
model pn = ger /dist=normal link=identity;
random animal/type=lin(1) LDATA=f5 solution;
parms (0.4764) (0.2189);
ods output solutionr = solutionrpncros;
output out=pncrospredito pred=pncrospredito;
run;


/* O código abaixo aplicado para a característica NO com
distribuição Poisson, referente aos dados do conjunto de
validação.*/


Title 'Analise NO - poisson-cross';
proc glimmix data=validacão noprofile IC=Q;
class animal ger;
model no = ger /dist=poisson link=log;
random animal p35 ipo /type=lin(1) LDATA=f5 solution ;
*output out=saidares resid=r pred=predicted;
parms (35);
ods output solutionr = blupnopoicros;
output out=nopoicrospredito pred=nopoicrospredito;
run;
```