A MODELAGEM DE UM ÍNDICE DE PRODUÇÃO CIENTÍFICA ATRAVÉS DE MODELOS LINEARES GENERALIZADOS HIERÁRQUICOS

MANOEL WALLACE ALVES RAMOS

Orientador: Prof[a] Dr[a] Maria Cristina Falcão Raposo

Co-orientador: Prof[a] Dr[a] Claudia Regina Oliveira de Paiva Lima

Área de Concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, janeiro de 2009

Ramos, Manoel Wallace Alves
A modelagem de um índice de produção científica através de modelos lineares generalizados hierárquicos / Manoel Wallace Alves Ramos - Recife : O Autor, 2009.
xi, 66 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CCEN. Estatística, 2009.

Inclui bibliografia e apêndice.

1. Estatística Aplicada – Modelagem. I. Título.

519.536 CDD (22.ed.) MEI2009-026

Universidade Federal de Pernambuco
Pós-Graduação em Estatística

20 de fevereiro de 2009
(data)

Nós recomendamos que a dissertação de mestrado de autoria de

Manoel Wallace Alves Ramos

intitulada

"A modelagem de um índice de produção científica através de Modelos Lineares Hierárquicos Generalizados"

seja aceita como cumprimento parcial dos requerimentos para o grau de Mestre em Estatística.

Klaus Leite Pinto Vasconcellos
Coordenador da Pós-Graduação em Estatística

Prof. Klaus Leite P. Vasconcellos
Coordenador de Pós Graduação em Estatística
UFPE

Banca Examinadora:

Maria Cristina Falcão Raposo
Maria Cristina Falcão Raposo orientador

Viviana Giampaoli
Viviana Giampaoli (USP)

Cristiano Ferraz
Cristiano Ferraz

Este documento será anexado à versão final da dissertação.

Até aqui tem me ajudado O Senhor.

Unicamente a Ele dedico este trabalho.

Agradecimentos

Primeiramente agradeço a Deus se não fosse por ele nada disso teria sido feito.

Aos meus pais, Marié e Manoel Cícero, pelo apoio, amor e incentivo em todos os momentos da minha vida.

A minha querida esposa Darlene pela constante paciência, apoio, amor e carinho.

Ao meu irmão Wanderson pela força e por sempre torcer por mim.

Às minhas orientadoras, Maria Cristina e Claudia Regina, pela confiança, dedicação, paciência e competência durante todo o período de desenvolvimento deste trabalho.

Agradeço aos professores do departamento de matemática, Manoel Lemos, César Castilho, Antônio Carlos e Adriano pedrosa, pelo grande incentivo.

Aos meus amigos do cursos de mestrado: Abrão, Alice, Andréa, Cícero, Fábio, Hemílio, Izabel, Juliana, Lídia, Marcelo, Olga, Raphael, Walmir, Wagner e Wilton. Vocês me ensinaram muitas coisas, me ajudaram em diversos momentos que precisei e tornaram essa trajetória muito mais fácil e divertida.

A Valéria Bittencourt pela competência e eficiência.

Agradeço aos professores da pós-graduação em estatística da UFPE pela oportunidade de melhorar meus conhecimentos e minha formação.

Aos participantes da banca examinadora, pelas sugestões.

À CAPES, pelo apoio financeiro.

Agradeço a todos que de forma direta ou indireta contribuíram para a conclusão deste trabalho.

Resumo

Nos últimos anos tem sido crescente o interesse de especialistas e autoridades governamentais por indicadores quantitativos da produção científica. A proposta desse trabalho é criar um índice de produção científica e construir um modelo que explique este índice. Para identificação do modelo mais adequado para atingir os objetivos foram ajustados modelos dentre os da classe dos modelos lineares generalizados hierárquicos(MLGH). Assim como nos modelos lineares generalizados(MLG), as variáveis-resposta dos MLGH podem seguir qualquer distribuição dentro da família exponencial. Neste trabalho a variável-resposta de interesse segue uma distribuição Poisson. Os dados utilizados se referem a 977 professores doutores da UFPE com dedicação exclusiva, sendo considerados 17 indicadores de produção nos anos de 2004 a 2006. As variáveis explicativas, consideradas como candidatas a explicar o índice no 1° nível hierárquico foram: idade, gênero e tempo de serviço. Considerando que os professores encontram-se agrupados nos centros acadêmicos, estes foram definidos como o 2° nível hierárquico e, como variáveis candidatas a fazer parte do modelo neste nível: o percentual de Bolsista de produtividade do CNPq e, área do conhecimento. Concluímos que a significativa diferença na produção científica ocorre nos centros acadêmicos cuja quantidade relativa de bolsistas de produtividade do CNPq é grande, com maior participação de docentes com idade até 50 anos. Em alguns centros as mulheres apresentam maior produção do que os homens, resultando diferenças estatísticas significativas por sexo e área do conhecimento.

Palavras-chave: modelo hierárquico, produção científica, indicador de produção, MLGH.

Abstract

In recent years, there is a growing interest of specialists and government authorities for quantitative indicators of scientific production. The purpose of this work is to create an index of scientific production and to build a model that explains this index. To identify the most appropriate model to achieve the objectives, models were adjusted among the class of hierarchical generalized linear models (HGLM). As in generalized linear models (GLM), HGLM response variables can follow any distribution within the exponential family. In this study the response variable of interest follows a Poisson distribution. The data used refer to 977 UFPE full professors, using 17 indicators of production from 2004 to 2006. The explanatory variables considered as essential to explain the index in the first level were age, gender and length of service. Considering that professors are grouped in academic centers, they were defined as the second level, and as essential variables to be part of the model at this level: the percentage of productivity(CNPq), and area of knowledge. We concluded that significant differences in scientific production occurs in the academic centers where there are many researchers younger than 50 years old. In some centers women have a higher production than men, resulting statistical differences by gender and area of knowledge.

Keywords: hierarchical model, scientific production, production indicator, HGLM.

Sumário

Lista de Figuras **xi**

Lista de Tabelas **xii**

1 Introdução **1**

1.1 Bibliometria 2

1.1.1 As Três Leis Clássicas 3

1.2 Cienciometria 5

1.3 Informetria 6

1.4 Webometria 7

1.5 Objetivo 8

1.6 Plataforma Computacional 9

2 Modelo Linear Generalizado **10**

2.1 Definição 10

2.2 A função *Deviance* 13

2.3 Modelo Normal e Modelo Poisson 15

2.3.1 Modelo Normal 15

2.3.2 Modelo Poisson 16

2.4 Estimação de $\underset{\sim}{\beta}$ 17

3 Modelo Linear Hierárquico **20**

3.1 O Modelo Linear Hierárquico com 2 Níveis 21

3.1.1 Anova com 1 Fator e Efeitos Aleatórios 22

3.1.2 Regressão de Médias como Respostas 24

3.1.3 Modelo de Regressão com Coeficientes Aleatórios 25

3.1.4 Interceptos e Inclinações como Respostas 27

3.1.5 Forma Geral do Modelo 29

3.2 Métodos de Estimação dos Parâmetros 29

3.3 Testes de Hipóteses 34

3.3.1 Teste de Hipóteses para Efeitos Fixos 34

3.3.2 Teste de Hipóteses para Efeitos Aleatórios 36

3.3.3 Teste de Hipóteses para Componentes de Variância e Covariância . 37

4 Modelos Lineares Generalizados Hierárquicos **39**

4.1 O Modelo Hierárquico Normal 40

4.2 O Modelo Hierárquico Poisson 41

4.3 Estimação em MLGH 43

5 Aplicação **45**

5.1 Introdução 45

5.2 Índice de Produção Científica 48

5.3 Modelagem Através do MLGH 53

5.3.1 Variáveis Explicativas 53

5.3.2 O Modelo 56

6 Conclusões e Sugestões para Trabalhos Futuros **60**

Referências Bibliográficas **61**

Apêndice **66**

Lista de Figuras

3.1 Estrutura dos dados para um modelo multinível com 2 níveis 22

5.1 Histograma do Índice de produção 52

Lista de Tabelas

2.1 Ligações canônicas de algumas distribuições 13

5.1 Pesos utilizados para a construção do índice de produção 50
5.2 Estatística descritiva do índice de produção por centro 53
5.3 Estatísticas descritivas das variáveis do modelo 57
5.4 Estimativas dos parâmetros do modelo 59

CAPÍTULO 1

Introdução

Nas últimas décadas, acompanhando a expansão da ciência e da tecnologia, tornou-se cada vez mais evidente a necessidade de mensurar tais avanços e de avaliar os desenvolvimentos alcançados pelas diversas áreas do conhecimento. Neste sentido, apontou-se para a medição das taxas de produtividade dos centros de pesquisa e dos investigadores individuais, para a detecção daquelas instituições e áreas com maiores potencialidades visando o estabelecimento das prioridades no momento da alocação de recursos públicos (Vanti, 2002).

Segundo Macias-Chapula(1998) os indicadores da atividade científica estão no centro dos debates, sob a perspectiva das relações entre o avanço da ciência e da tecnologia, por um lado, e o progresso econômico e social, por outro. Revisões de políticas científicas pareceriam inconcebíveis, hoje, sem recorrer aos indicadores existentes. Se por muito tempo o foco das avaliações permaneceu orientado para medir os insumos, como verbas e pessoal de P&D (pesquisa e desenvolvimento), crescentemente o interesse está se voltando para os indicadores de resultados (Okubo, 1997). Em tudo o que se refere à ciência, os indicadores bibliométricos e cienciométricos tornaram-se essenciais.

Diversas formas de medição voltadas para avaliar a ciência e o fluxo de informação foram criadas. Entre elas se destacam a bibliometria, a cienciometria, a informetria e a mais recente delas a webometria. Apesar destas formas de medição apresentarem algumas semelhanças ou pontos de convergência, possuem enfoques e funções diferentes.

1.1 Bibliometria

De acordo com Tangue-Sutcliffe(1992) Bibliometria é:

"[...] o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. A bibliometria desenvolve padrões e modelos matemáticos para medir esses processos, usando seus resultados para elaborar previsões e apoiar tomadas de decisão ".

De acordo com Vanti(2002) há controvérsias com respeito a origem da bibliometria, embora Lawani(1981) e Sengupta(1992) acreditem que o termo bibliometria tenha sido cunhado por Alan Pritchard em 1969. Fonseca(1973) refere que foi Paul Otlet que utilizou a palavra bibliometria pela primeira vez em sua obra intitulada *Traité de documentatión*, de 1934 e que Pritchard popularizou a palavra "bibliometria" a partir de um artigo que discutia a polêmica, "bibliografia estatística ou bibliometria ?" quando sugeriu que esta última deveria substituir o termo "bibliografia estatística", que vinha sendo utilizado desde a citação feita em 1922 por Edward Wyndham Hulme. De acordo com Nicholas e Ritchie(1978) a diferença básica entre a tradicional bibliografia e a bibliometria é que esta última utiliza mais métodos quantitativos do que discursivos.

Inicialmente a bibliometria estava voltada para a medida de livros (quantidade de edições e exemplares, quantidade de palavras contidas nos livros, espaço ocupado pelos livros nas bibliotecas, estatísticas relativas à indústria do livro), aos poucos foi se voltando para o estudo de outros formatos de produção bibliográfica, tais como artigos em periódicos e outros tipos de documentos, para depois ocupar-se, também, da produtividade de

autores e do estudo de citações. A bibliometria desenvolve-se basicamente a partir da elaboração de leis empíricas sobre o comportamento da literatura, de autoria de Lotka, Bradford & Zipf (Araújo, 2006).

1.1.1 As Três Leis Clássicas

- *Lotka*

A Lei de Lotka, ou Lei do Quadrado Inverso, foi formulada em 1926. Lotka observou num estudo sobre produtividade dos cientistas a partir da contagem de autores presentes no *Chemical Abstracts*, entre 1909 e 1916, que uma grande proporção da literatura científica era produzida por um pequeno número de autores e um grande número de pequenos produtores se igualava, em produção, ao reduzido número de grandes produtores. A partir daí formulou a lei dos quadrados inversos, segundo a qual o número de autores que tem n contribuições em um determinado campo científico é aproximadamente $\frac{1}{n^2}$ daqueles que só fazem uma contribuição e que a proporção daqueles que fazem uma única contribuição é de mais ou menos 60%. Ou seja, 60% dos autores teriam apenas uma publicação, enquanto 15% teriam duas e 7% teriam 3 e assim por diante.

A Lei de Lotka foi, desde então, objeto de larga produção científica. Muitos autores, desde que a lei foi formulada em 1926, produziram trabalhos criticando, replicando e/ou reformulando esta lei bibliométrica, (Araújo, 2006).

- *Bradford*

A Lei de Bradford, relacionada à dispersão da literatura periódica científica, enuncia que "*se periódicos científicos forem ordenados em ordem decrescente de produtividade de artigos sobre determinado assunto, poderiam ser divididos em um núcleo de periódicos mais particularmente dedicados ao assunto e em vários grupos ou zonas, contendo o mesmo número de artigos que o núcleo. O número de periódicos* (n)*, no núcleo e zonas subsequentes, variaria na proporção* $1 : n : n^2$ *[...]*"(Brookes, 1990). Bradford chegou a

essa conclusão em seus estudos em 1934. Ele percebeu que em uma coleção de periódicos sobre geofísica, existe sempre um núcleo menor de periódicos relacionados de maneira próxima ao assunto e um número maior de periódicos relacionados de maneira estreita, sendo que o número de periódicos de cada zona aumenta, enquanto a produtividade diminui. Analisando 326 periódicos, ele descobriu que 9 periódicos continham 429 artigos, 59 continha 499 e 258 continham 404 artigos. Dessa forma ordenando uma grande coleção de periódicos em ordem de produtividade decrescente relevante a um dado assunto, três zonas aparecem, cada uma contendo 1/3 do total de artigos relevantes (a primeira zona contém um pequeno número de periódicos altamente produtivos, a segunda contém um número maior de periódicos menos produtivos, e a terceira inclui mais periódicos que a segunda zona, mas, cada periódico com menos produtividade). Bradford viu que era por essa razão que os índices tinham dificuldade para atingir a cobertura completa de assuntos. Havendo grande número de periódicos na zona exterior, Bradford constatou que mais da metade do total de artigos úteis não estavam sendo cobertos pelos serviços de indexação e resumos (Araújo, 2006).

A Lei de Bradford é um instrumento útil para o desenvolvimento de políticas de aquisição e de descarte de periódicos, ligadas a gestão de sistemas de recuperação da informação, gestão da informação e do conhecimento científico e tecnológico. É possível estimar a magnitude de determinada área bibliográfica e o custo de toda e qualquer fração específica da bibliografia (Guedes & Borschiver, 2005).

- *Zipf*

A terceira das leis bibliométricas clássicas é a Lei de Zipf, também conhecida como lei do mínimo esforço, formulada em 1949 e que descreve a relação entre palavras num determinado texto suficientemente grande e a ordem de série destas palavras (contagem de palavras em largas amostragens). Zipf observou que, num texto suficientemente longo, existia uma relação entre a frequência que uma dada palavra ocorria e sua posição na

lista de palavras ordenadas segundo sua frequência de ocorrência. Essa lista era confeccionada levando-se em conta a frequência decrescente de ocorrências. A posição nesta lista dá-se o nome de ordem de série (rank). Assim, a palavra de maior frequência de ocorrência tem ordem de série 1, a de segunda maior frequência de ocorrência, ordem de série 2 e, assim, sucessivamente. Zipf observou, também, que o produto da ordem de série (r) de uma palavra, pela sua frequência de ocorrência (f) era aproximadamente constante (k). Enunciou assim que $r.f = k$, o que ficou conhecido como Primeira Lei de Zipf.

Esta Lei de Zipf é muito utilizada para indexar artigos científicos. De acordo com a Lei de Zipf pode-se medir a frequência de aparecimento de diversas palavras em vários textos objetivando gerar uma lista de termos de uma determinada disciplina. De acordo com Zipf, em certas disciplinas determinadas palavras têm probabilidade maior de ocorrência, enquanto que algumas têm menor frequência, e outras são raramente utilizadas, (Guedes e Borschiver, 2005).

1.2 Cienciometria

Segundo Vanti(2002) o termo cienciometria surgiu na antiga União Soviética e Europa Oriental e foi empregado especialmente na Hungria. Dobrov & Karennoi foram os primeiros autores a utilizar o termo em uma publicação do All-Union Institute for Scientific and Technical Information (VINITI). Originalmente o termo referia-se a aplicação de métodos quantitativos para o estudo da história da ciência e do progresso tecnológico. A cienciometria foi definida inicialmente como "a medição do processo informático", onde o termo "informático" significava "a disciplina do conhecimento que estudava a estrutura e as propriedades da informação científica e as leis do processo de comunicação" (Spinak, 1996). Tal termo ficou bastante conhecido com o início da publicação, em 1977, da revista *Scientometrics*, editada originalmente na Hungria e atualmente na Holanda (Vanti, 2002).

Macias-Chapula(1998) define cienciometria como:

"[...]estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. A cienciometria é um segmento da sociologia da ciência, sendo aplicada no desenvolvimento de políticas científicas. Envolve estudos quantitativos das atividades científicas, incluindo a publicação e, portanto, sobrepondo-se à bibliometria".

Para Van Raan(1997) a cienciometria realiza estudos quantitativos em ciência e tecnologia e tenta descobrir os laços existentes entre ambas, visando o avanço do conhecimento e buscando relacioná-lo com questões sociais e de políticas públicas.

Fazendo uma distinção entre bibliometria e cienciometria, Spinak(1998) afirma que:

"La bibliometría estudia la organización de los sectores científicos y tecnológicos a partir de las fuentes bibliográficas y patentes para identificar los actores, sus relaciones y sus tendencias. Por el contrario, la cienciometría trata con las varias mediciones de la literatura, de los documentos y otros medios de comunicación, mientras que la bibliometría tiene que ver con la produtividad y utilidad científica".

"La cienciometría aplica técnicas bibliométricas a la ciencia [...] pero va mas allá de las técnicas bibliométricas, pues también examina el desarollo y las políticas científicas. [...] la cienciometría puede establecer comparaciones entre las políticas de investigación entre los países analizando sus aspectos económicos y sociales".

1.3 Informetria

Para Macias-Chapula(1998) informetria é:

"[...]estudo dos aspectos quantitativos da informação em qualquer formato, e não apenas registros catalográficos ou bibliografias, referente a qualquer grupo social, e não apenas aos cientistas. A informetria pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites tanto da bibliometria como da cien-

ciometria".

Considerada a mais completa dos três indicadores, a informetria engloba a bibliometria e a cienciometria. O termo foi proposto pela primeira vez na Alemanha, em 1979, por Otto Nacke e aceito definitivamente em 1989 no Encontro Internacional de Bibliometria, que passou a ser chamado de Conferência Internacional de Bibliometria, Cienciometria e Informetria (Araújo, 2006). A informetria tem como prioridade o desenvolvimento de modelos matemáticos e, em segundo lugar, a determinação de medidas para o fenômeno estudado. Os modelos oferecem uma base prática para a tomada de decisões, e seu valor está na sua capacidade de sintetizar, em poucos parâmetros, as características de muitos grupos de dados: formato completo, concentração, difusão e mudança através do tempo. A informetria é bastante útil na administração de coleções em bibliotecas, no desenvolvimento de políticas científicas e pode ajudar na tomada de decisões em relação ao desenho e manutenção de sistemas de recuperação

A informetria se distingue da cienciometria e da bibliometria no que diz respeito ao universo de objetos e sujeitos que estuda, não se limitando apenas à informação registrada, dado que pode analisar também os processos de comunicação informal, inclusive falada, e dedicar-se a pesquisar os usos e necessidades de informação dos grupos sociais desfavorecidos, e não só das elites intelectuais (Vanti, 2002).

1.4 Webometria

A webometria, definida como o uso de técnicas bibliométricas à World Wide Web (WWW), é um sistema de estudos de relacionamento de diferentes sites na rede. Essa técnica também pode ser usada para mapear (chamada de "scientifc mapping" na pesquisa bibliométrica tradicional) áreas da Web que se tornaram mais usadas, baseadas no número de vezes que foram conectados por outros websites.

Como aplicação de métodos informétricos na WWW, pode-se afirmar que a webome-

tria é uma forma de reconhecimento da importância da rede como meio de informação e comunicação para a ciência e a academia, setores aos quais os estudos quantitativos têm servido.

Segundo Vanti(2002) o termo cunhado por Almind e Ingwersen em 1997, consiste na aplicação de métodos informétricos à Word Wide Web (Web ou www) para fins de medir seu fluxo.

Dentre as medições que podem ser realizadas no campo da webometria, destacam-se a frequência de distribuição e as classificações que compreendem categorias tais como homepages pessoais, institucionais ou organizacionais. Podem-se também realizar mensurações em tempos diferentes para comparar a evolução de uma instituição ou país na rede, para calcular o tamanho médio de uma página expressado em *bytes*, o número médio em *links* por página e a densidade média por *link*. Outros tipos de análise referem-se às citações entre páginas, conhecidas como *links*, *hyperlinks* ou *weblinks*, e estes são vistos como indicadores da importância global de um site ou um espaço (Vanti, 2002).

1.5 Objetivo

Dentre as diversas formas de medições voltadas para avaliar a ciência, o presente trabalho se enquadra no tipo informetria, pois tem como objetivos:

- propor um índice de produção científica;
- medir a produção científica da UFPE a partir do índice proposto;
- identificar as possíveis variáveis explicativas da produção docente;
- encontrar um modelo matemático que explique a produção docente.

Para identificação do modelo matemático mais adequado para atingir os objetivos foram ajustados modelos dentre os da classe dos modelos lineares hierárquicos generalizados, devido a natureza dos dados.

Um breve resumo da teoria dos modelos lineares generalizados, dos modelos lineares hierárquicos e dos modelos lineares generalizados hierárquicos encontra-se apresentado nos capítulos 2, 3 e 4, respectivamente, desta dissertação. No capítulo 5 encontra-se o resultado da análise dos dados da produção científica dos professores da UFPE no período de 2004 a 2006 e no capítulo 6 as conclusões e sugestões para trabalhos futuros.

1.6 Plataforma Computacional

Os resultados numéricos apresentados nesta dissertação foram obtidos utilizando o ambiente de programação e análise de dados R em sua versão 2.6.0 para sistema operacional Microsoft Windows. O R se encontra disponível gratuitamente através do site `http://www.R-project.org`.

A presente dissertação de mestrado foi digitada utilizando o sistema de tipografia LaTeX, que consiste em uma série de macros ou rotinas do sistema TeX que facilitam o desenvolvimento da edição do texto. Detalhes sobre o sistema de tipografia LaTeX podem ser encontrados no site `http://www.latex-project.org/`.

CAPÍTULO 2

Modelo Linear Generalizado

A unidade de muitos métodos estatísticos é evocada por um modelo linear generalizado, à qual está vinculada a idéia de uma família exponencial de distribuições de probabilidades associadas a uma variável aleatória. Este capítulo aborda a definição do modelo linear generalizado, o estudo da função "Deviance", o modelo linear generalizado com distribuições Normal e Poisson, uma vez que o modelo normal é o mais conhecido e o modelo Poisson que é o modelo utilizado na nossa aplicação. Por fim é apresentado o algoritmo para a estimação dos parâmetros do modelo.

2.1 Definição

Esta classe de modelos é uma extensão de modelos lineares clássicos e foi desenvolvida por Nelder e Wedderburn(1972), e permite tratar de uma mesma forma, uma grande quantidade de modelos conhecidos e largamente aplicados. Nelder e Wedderburn(1972) mostraram que uma série de técnicas comumente estudadas separadamente podem ser reunidas sob o nome de Modelos Lineares Generalizados(MLG). Assim são casos especiais de MLG:

- Modelos clássicos de Regressão com erro normal;

- Modelos clássicos de análise de variância e de covariância com erro normal;
- Modelo de análise de variância com erros aleatórios;
- Modelo log-linear aplicado à análise de tabelas de contingência;
- Modelos logit e probit para análise de proporções, dentre outros.

Um modelo linear generalizado é definido a partir de duas componentes mais uma função relacionando estas componentes.

1. *Componente aleatória*

   Definida a partir das variáveis respostas $Y_1 \dots Y_n$, supostamente independentes, cada uma com densidade na forma da família exponencial dada por:

   $$f(y; \theta_i, \phi) = exp[\phi\{y\theta_i - b(\theta_i)\} + c(y, \phi)], \tag{2.1}$$

   em que $b(.)$ e $c(.)$ são funções supostas conhecidas, θ é o parâmetro natural que caracteriza a densidade e ϕ^{-1} é o parâmetro de dispersão. A esperança e a variância da variável aleatória Y são dadas por:

   $$E(Y_i) = \mu_i = b'(\theta_i),$$

   $$Var(Y_i) = \phi^{-1} V_i,$$

   $V = d\mu/d\theta$ é a função de variância.

   Esta família exponencial representa a única fonte de variação aleatória do modelo linear generalizado.

2. *Componente sistemática*

   É representada como uma função linear de um conjunto de parâmetros desconhecidos. Dessa forma temos:

$\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{ip})'$ representa os valores de p variáveis explicativas, $\underset{\sim}{\beta} = (\beta_1, \ldots, \beta_p)'$, $p < n$ é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{X} = (\underset{\sim}{x}'_1, \underset{\sim}{x}'_2 \ldots, \underset{\sim}{x}'_n)'$ é a matriz do modelo e $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \underset{\sim}{x}'_i\underset{\sim}{\beta}$ ou $\underset{\sim}{\eta} = \mathbf{X}\underset{\sim}{\beta}$ é o vetor de preditores lineares que corresponde a componente sistemática, e descreve a parte determinística do modelo. Em geral, a relação entre a componente sistemática e a componente aleatória não é uma relação de adição.

3. *Função de ligação*

   É uma função g, contínua e diferenciável, que relaciona o preditor linear (η) e o valor esperado (μ) de Y. Ou seja:

$$g(\mu_i) = \eta_i. \tag{2.2}$$

No modelo linear clássico tem-se $\eta = \mu$ que é chamada ligação identidade. Essa ligação é adequada no sentido em que ambos η e μ podem assumir valores na reta real conforme referem McCullagh e Nelder(1989). Entretanto, certas restrições surgem quando se trabalha por exemplo, com a distribuição de Poisson em que $\mu > 0$ e portanto, a ligação identidade não deve ser usada, pois η poderá assumir valores negativos dependendo dos valores obtidos para $\underset{\sim}{\widehat{\beta}}$. Além disso, dados de contagem dispostos em tabelas de contingência sob a suposição de independência levam naturalmente a efeitos multiplicativos cuja linearização pode ser obtida através da ligação logarítmica, isto é, $\eta = log\mu$ de onde se tem $\mu = e^{\eta}$ (Demétrio, 1993).

As funções de ligação mais utilizadas para os casos em que as variáveis respostas tem distribuição binomial são:

$$\eta = ln\frac{\mu}{1-\mu} \qquad Logit$$

$$\eta = \phi^{-1}(\mu) \qquad Probit$$

$$\eta = ln[-ln(1-\mu)] \qquad Complemento\ log - log$$

Uma família de ligações importantes, principalmente para dados com média positiva é a família potência. Essa família, pode ser especificada por:

$$\eta = \begin{cases} \frac{\mu^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ ln\mu, & \lambda = 0 \end{cases}$$

ou então,

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

A função de ligação é chamada canônica, quando o parâmetro natural coincide com o preditor linear, ou seja, $\theta = \eta$.

A função de ligação canônica fornece, para o modelo, estatísticas suficientes de dimensão mínima igual a p. Deve ser lembrado, porém, que embora as ligações canônicas levem as propriedades estatísticas desejáveis para o modelo, principalmente no caso de amostras pequenas, não há nenhuma razão *a priori* para que os efeitos sistemáticos do modelo devam ser aditivos na escala dada por tais ligações (McCullagh e Nelder, 1989).

A Tabela 2.1 mostra as ligações canônicas de algumas distribuições.

Distribuição	Ligação canônica
Normal	$\eta = \mu$
Poisson	$\eta = ln\mu$
Binomial	$\eta = ln\frac{\mu}{1-\mu}$
Binomial Negativa	$\eta = ln\frac{\mu}{\mu+\kappa}$
Gama	$\eta = \frac{1}{\mu}$
Normal Inversa	$\eta = \frac{1}{\mu^2}$

Tabela 2.1: Ligações canônicas de algumas distribuições

2.2 A função *Deviance*

Considerando Y uma variável aleatória com função de probabilidade $f(y;\theta_i,\phi)$ da família exponencial de distribuições com um parâmetro desconhecido, o logaritmo da

função de verossimilhança para a observação y é dado por:

$$L(\mu; y) = \phi\{y\theta_i - b(\theta_i)\} + c(y, \phi).$$

Considerando $Y_1 \ldots Y_n$, n observações independentes de Y, temos o logaritmo da função de verossimilhança para todas as amostras como sendo

$$L(\underset{\sim}{\mu}; \underset{\sim}{y}) = \sum_{i=1}^{n} L(\mu_i; y_i),$$

em que $\underset{\sim}{y} = (y_1, \ldots, y_n)'$ e $\underset{\sim}{\mu} = (\mu_1, \ldots, \mu_n)'$.

Para um conjunto de dados podem ser considerados, dois modelos extremos, um modelo "minimal" que contém o menor conjunto de termos que o problema permite, e um modelo "saturado" ($p = n$) em que todos os $\mu's$ são diferentes, correspondendo a um parâmetro para cada observação. O modelo saturado fornece uma base para medir a discrepância em relação a um modelo intermediário com p parâmetros. Na prática deseja-se encontrar um modelo com p parâmetros linearmente independentes entre esses dois, o chamado "modelo sob investigação". Em cada estágio, troca-se crescentemente a medida de bondade do ajuste para o modelo sob investigação, associado a uma crescente complexidade do modelo. Uma medida desta discrepância, foi definida por Nelder e Wedderburn(1972), denominada *deviance*, e é dada por:

$$D^*(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}) = \phi D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}) = 2\{L(\underset{\sim}{y}; \underset{\sim}{y}) - L(\underset{\sim}{\hat{\mu}}; \underset{\sim}{y})\},$$

que é uma distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança $\underset{\sim}{\hat{\beta}}$, onde $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$ em que $\hat{\eta}_i = \underset{\sim}{x}_i'\underset{\sim}{\hat{\beta}}$.

Temos que a função $D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}})$ pode ser escrita alternativamente por:

$$D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}) = 2\sum_{i=1}^{n}\{y_i(\hat{\theta}_i^0 - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\hat{\theta}_i^0))\},$$

denotando por $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$ e $\hat{\theta}_i^0 = \theta_i(\hat{\mu}_i^0)$ as estimativas de máxima verossimilhança de θ_i para os modelos com p parâmetros ($p < n$) e saturado ($p = n$), respectivamente.

2.3 Modelo Normal e Modelo Poisson

Nesta seção estão apresentadas as partes aleatória e sistemática, a função de ligação e a função "deviance" dos modelos Normal e Poisson.

2.3.1 Modelo Normal

Abordaremos o modelo normal, uma vez que a distribuição normal é a mais familiar das distribuições de probabilidade e de fácil entendimento.

A definição desse modelo, como um MLG, pode ser formalizada conforme segue:

1. *Componente aleatória*

   Seja Y uma variável aleatória com distribuição normal de média μ e variância σ^2, $Y \sim N(\mu, \sigma^2)$. Temos que a função densidade de probabilidade de Y é dada por:

   $$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\},$$

   em que $-\infty < y < \infty$, $-\infty < \mu < \infty$ e $\sigma^2 \geq 0$.

   Escrevendo a função densidade de probabilidade de Y como um membro da família exponencial de distribuições temos

   $$f_Y(y) = exp[\{\frac{1}{\sigma^2}(\mu y - \frac{\mu^2}{2}) - \frac{1}{2}\{log 2\pi\sigma^2 + \frac{y^2}{\sigma^2}\}],$$

   onde $\theta = \mu$, $b(\theta) = \theta^2/2$, $\phi = \sigma^{-2}$ e $c(y, \phi) = \frac{1}{2} log\phi/2\pi - \frac{\phi y^2}{2}$ obtém -se a expressão (2.1). Podendo-se verificar que:

   $$E(y) = \mu = \frac{db(\theta)}{d\theta} = \theta, \qquad V(Y) = \phi^{-1}\frac{d\mu}{d\theta} = \sigma^2 \quad e \quad V = 1.$$

2. *Componente sistemática*

   A componente sistemática é a estrutura linear $\eta = \underset{\sim}{x}_i'\underset{\sim}{\beta}$.

3. *Função de Ligação e Deviance*

   A função de ligação canônica é dada por $\eta = \mu$.

   A função "deviance" para o modelo normal, onde $\theta_i = \mu_i$, $\hat{\theta}_i^0 = y_i$ e $\hat{\theta}_i = \hat{\mu}_i$, é dada por:

$$D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}) = 2\sum_{i=1}^{n}\{y_i(y_i - \hat{\mu_i}) + \frac{\hat{\mu}_i^2}{2} - \frac{y_i^2}{2}\} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2,$$

   que coincide com a soma dos quadrados dos resíduos.

2.3.2 Modelo Poisson

A definição do modelo Poisson, como um MLG, pode ser formalizada conforme segue:

1. *Componente aleatória*

   Seja $Y \sim P(\mu)$, então temos

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!},$$

   onde $y = 0, 1, \ldots$.

   A função densidade de Y pode ser escrita como um membro da família exponencial de distribuições conforme a expressão (2.1), da seguinte forma:

$$f_Y(y) = exp\{ylog\mu - \mu - log(y!)\},$$

   onde $log\mu = \theta$, $b(\theta) = e^\theta$, ϕ =1 e $c(y, \phi) = -log(y!)$. Segue portanto que

$$E(Y) = \mu = \frac{db(\theta)}{d\theta}, \qquad V(Y) = \phi^{-1}\frac{d\mu}{d\theta} = \mu \quad e \quad V = \mu.$$

2. *Componente sistemática*

   Será a estrutura linear já definida $\eta = \underset{\sim}{x}_i'\underset{\sim}{\beta}$.

3. *Função de Ligação e Deviance*

Para o modelo Poisson, as funções de ligação mais utilizadas são:

- $\eta = log(\mu)$ que é a função de ligação canônica;
- $\eta = 3\mu^{1/3}$ função de ligação normalizadora e
- $\eta = 3\mu^{1/2}$ função de ligação estabilizadora.

A "Deviance" do modelo Poisson é dada por:

$$D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}}) = 2\sum_{i=1}^{n}\{y_i log(\frac{y_i}{\hat{\mu}_i}) - (y_i - \hat{\mu}_i)\},$$

tem-se que $\theta_i = log\mu_i$, o que implica em $\hat{\theta}_i^0 = logy_i$ e $\hat{\theta}_i = log\hat{\mu}_i$.

Se $y_i = 0$ o i-ésimo termo de $D(\underset{\sim}{y}, \underset{\sim}{\hat{\mu}})$ será $2\hat{\mu}_i$.

2.4 Estimação de $\underset{\sim}{\beta}$

Para o Modelo Linear Generalizado, as equações de máxima verossimilhança são não lineares, com exceção do modelo normal com ligação identidade e, logo, não podem ser resolvidas explicitamente. Métodos numéricos iterativos, como os de Newton-Rapson, equivalentes a técnica de escore de Fisher, são necessários para a obtenção de $\underset{\sim}{\hat{\beta}}$.

O logaritmo da função de verossimilhança de um MLG definido por (2.1) e (2.2) pode ser expresso na forma:

$$L(\underset{\sim}{\beta}; \underset{\sim}{y}) = \sum_{i=1}^{n}[\phi\{y\theta_i - b(\theta_i)\} + c(y_i, \phi)].$$

Assim a função escore total e a matriz de informação total de Fisher para $\underset{\sim}{\beta}$, quando ϕ é conhecido, são dadas por:

$$\mathbf{U}(\underset{\sim}{\beta}) = \frac{\partial L(\underset{\sim}{\beta}; \underset{\sim}{y})}{\partial \underset{\sim}{\beta}} = \phi \mathbf{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\underset{\sim}{y} - \underset{\sim}{\mu}),$$

$$\mathbf{K}(\underset{\sim}{\beta}) = E\left\{-\frac{\partial^2 L(\underset{\sim}{\beta};\underset{\sim}{y})}{\partial\underset{\sim}{\beta}\partial\underset{\sim}{\beta}'}\right\} = \phi\mathbf{X}'\mathbf{W}\mathbf{X},$$

respectivamente em que $\mathbf{X}$ é a matriz modelo $n\times p$ de posto completo, $\mathbf{W} = diag(w_1,\ldots,w_n)$ é a matriz de pesos, com $w_i = \left(\frac{d\mu_i}{d\eta_i}\right)^2 \frac{1}{V_i}$ e $\mathbf{V} = diag(V_1,\ldots,V_n)$, com $V_i = \frac{d\mu_i}{d\theta_i}$.

O processo iterativo de Newton-Raphson para a obtenção da estimativa de máxima verossimilhança de $\underset{\sim}{\beta}$, consiste em expandir a função escore $\mathbf{U}(\underset{\sim}{\beta})$ em torno de um valor inicial $\underset{\sim}{\beta}^{(0)}$, tal que

$$\mathbf{U}(\underset{\sim}{\beta}) \cong \mathbf{U}(\underset{\sim}{\beta}^{(0)}) + \mathbf{U}'(\underset{\sim}{\beta}^{(0)})(\underset{\sim}{\beta} - \underset{\sim}{\beta}^{(0)}),$$

onde $\mathbf{U}'(\underset{\sim}{\beta})$ é a primeira derivada de $U(\underset{\sim}{\beta})$ com respeito a $\underset{\sim}{\beta}$. Então se repetirmos o procedimento acima, teremos o processo iterativo

$$\underset{\sim}{\beta}^{(m+1)} = \underset{\sim}{\beta}^{(m)} + \{-\mathbf{U}'(\underset{\sim}{\beta}^{(m)})\}^{-1}\mathbf{U}(\underset{\sim}{\beta}^{(m)}),$$

$m = 0, 1, \ldots$. Como a matriz $-\mathbf{U}'(\underset{\sim}{\beta})$ pode não ser positiva definida, a aplicação do método de *escore* de Fisher substituindo a matriz $-\mathbf{U}'(\underset{\sim}{\beta})$ pelo correspondente valor esperado, pode ser o mais conveniente (Paula, 2004). Isso resulta no seguinte processo iterativo:

$$\underset{\sim}{\beta}^{(m+1)} = \underset{\sim}{\beta}^{(m)} + \mathbf{K}^{-1}(\underset{\sim}{\beta}^{(m)})\mathbf{U}(\underset{\sim}{\beta}^{(m)}),$$

$m = 0, \ldots$. Depois de algumas manipulações algébricas do lado direito da expressão acima, obtém-se um processo iterativo de mínimos quadrados reponderados

$$\underset{\sim}{\beta}^{(m+1)} = (\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(m)}\underset{\sim}{z}^{(m)}, \tag{2.3}$$

$m = 0, 1, \ldots$, em que $\underset{\sim}{z} = \underset{\sim}{\eta} + \mathbf{W}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}(\underset{\sim}{y} - \underset{\sim}{\mu})$. Observa-se que $\underset{\sim}{z}$ desempenha o papel de uma variável dependente modificada, enquanto $\mathbf{W}$ é uma matriz de pesos que muda a cada passo do processo iterativo. A convergência de (2.3) ocorre em um número finito de

passos, independente dos valores iniciais utilizados. É usual iniciar o processo iterativo com $\underset{\sim}{\eta}^{(0)} = g(\underset{\sim}{y})$. A estimação do parâmetro ϕ, quando o mesmo é desconhecido, pode ser vista em (Cordeiro e McCullagh, 1991).

CAPÍTULO 3

Modelo Linear Hierárquico

Na prática é muito comum encontrar dados que têm uma estrutura hierárquica ou de grupos. Estes tipos de dados podem ser estudados através dos chamados *modelos lineares hierárquicos* (MLH)(Goldstein, 1999). Estes modelos permitem que cada um dos níveis desta estrutura hierárquica seja especificado separadamente e que posteriormente sejam reunidos em um único modelo. Além disso permite a incorporação de efeitos aleatórios associados a cada um dos seus níveis de hierarquia.

Uma situação muito citada na literatura é o estudo da habilidade (ou proficiência) de estudantes. Neste caso temos então, um grupo de estudantes associados a escolas, ou seja, o nível 1 é o nível dos alunos e o nível 2 é o nível das escolas.

As idéias aqui expostas de forma introdutória estão amplamente desenvolvidas em Bryk & Raudenbush(2002) que, conjuntamente com Goldstein(1999), são consideradas como referências básicas em MLH. Entretanto, a denominação de "Modelo Linear Hierárquico" é bem mais antiga e, de acordo com Natis(2000), ela surgiu originalmente como fruto dos trabalhos de Lindley & Smith(1972) no desenvolvimento de métodos de estimação Bayesiana em modelos lineares.

Neste capítulo será apresentado uma introdução sobre os modelos lineares hierárquicos

com 2 níveis e alguns de seus principais submodelos. Na Seção 3.2 são abordados alguns métodos para a estimação dos parâmetros do modelo e na Seção 3.3 são descritos os testes de hipóteses para os efeitos fixos, para os efeitos aleatórios e para os componentes de variância e covariância do modelo.

3.1 O Modelo Linear Hierárquico com 2 Níveis

O modelo de regressão linear hierárquico ou multinível com dois níveis assume que há um conjunto de dados hierárquicos, que possui uma variável resposta (Y) que é medida no nível individual, e variáveis explicativas que podem residir no nível do indivíduo(X) e/ou do grupo (W), que é um nível mais elevado. O modelo pode ser visto como um sistema hierárquico de equações de regressão.

No modelo de regressão hierárquico temos que o subscrito i denota o i-ésimo indivíduo ($i = 1, \ldots, n_j$) do grupo j ($j = 1, \ldots, J$), ou seja ocorrem n_j unidades do nível 1 (*do indivíduo*) para cada unidade j ($j = 1, \ldots, J$) do nível 2 (*do grupo*). A Figura 3.1 mostra um esquema dos dados estruturados segundo um modelo hierárquico com dois níveis.

Apresentaremos o modelo de regressão hierárquica começando com alguns modelos particulares até obter a forma geral do modelo.

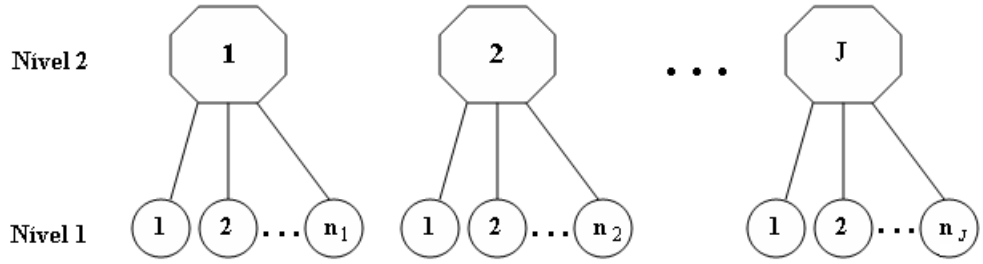

Figura 3.1: Estrutura dos dados para um modelo multinível com 2 níveis

3.1.1 Anova com 1 Fator e Efeitos Aleatórios

Segundo Bryk & Raudenbush(2002) temos o modelo linear hierárquico mais simples, quando não existem variáveis explicativas em nenhum dos dois níveis. É o modelo com 1 fator com efeitos aleatórios.

Temos então o modelo do nível 1

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{3.1}$$

com $i = 1, 2, \ldots, n_j$ e $j = 1, 2, \ldots, J$ onde

Y_{ij} é a variável resposta do i-ésimo indivíduo do nível 1 para o j-ésimo grupo do nível 2;

β_{0j} é a resposta esperada para o j-ésimo grupo;

e_{ij} é o erro aleatório associado à i-ésima unidade do nível 1 agrupado na j-ésima unidade do nível 2, com $e_{ij} \sim N(0, \sigma^2)$ e $e'_{ij}s$ independentes.

E o modelo do nível 2 é dado por:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.2}$$

onde

γ_{00} é a média da variável resposta para a população;

u_{0j} é o efeito aleatório associado ao j-ésimo grupo, $u_{0j} \sim N(0, \tau_{00})$, com $u'_{0j}s$ independentes entre si e $u'_{0j}s$ independentes de $e'_{ij}s$.

Quando substituimos a equação (3.2) na equação (3.1) obtemos o modelo combinado

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{3.3}$$

A variância da resposta é dada por:

$$Var(Y_{ij}) = Var(\gamma_{00} + u_{0j} + e_{ij}) = \tau_{00} + \sigma^2.$$

O modelo hierárquico 3.3 é chamado *totalmente não condicional*, pois tanto o nível 1 quanto o nível 2 não possuem nenhum preditor. O modelo é considerado de efeitos aleatórios, pois os efeitos dos grupos (u_{0j}) são interpretados como aleatórios. A variância de Y_{ij} é decomposta em duas componentes independentes: σ^2 que é a variância dos erros do nível 1 (do indivíduo), aqui denominado e_{ij}; e τ_{00} que é a variância dos erros do nível 2 (do grupo), definidos por u_{0j}.

Um parâmetro de grande utilidade que está associado à ANOVA com 1 fator e efeitos aleatórios é o *coeficiente de correlação intra-classe*, dado por:

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \tag{3.4}$$

Ele representa a proporção da variância da resposta explicada pela variabilidade entre as unidades do nível 2.

3.1.2 Regressão de Médias como Respostas

Neste modelo são incorporadas variáveis explicativas no nível 2, buscando explicar a variabilidade dos coeficientes β_{0j} entre as unidades do nível 2.

Temos que o modelo do nível 1 definido em (3.1) é igual ao caso da ANOVA com um fator e efeitos aleatórios, ou seja, as equações para o nível 1 e nível 2 são respectivamente:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + e_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \end{aligned} \tag{3.5}$$

com $i = 1, 2, \ldots, n_j$ e $j = 1, 2, \ldots, J$ onde

β_{0j} é o valor esperado da variável resposta de um modelo de regressão linear onde as variáveis explicativas correspondem a característica do grupo j. E, nesse caso temos a variável explicativa (W) para o nível 2.

Substituindo a equação (3.5) na equação (3.1) obtemos o modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{oj} + e_{ij} \tag{3.6}$$

onde

γ_{00} é o intercepto médio dos grupos para W_j igual a zero;

γ_{01} é a diferença média entre os J grupos;

u_{0j} é o efeito aleatório do j-ésimo grupo sobre o intercepto para W_j igual a zero;

e e_{ij} é definido como no item 3.1.1.

O coeficiente ρ apresentado na equação (3.4) agora é chamado coeficiente de correlação intra-classe condicional e continua representando o grau de dependência entre indivíduos de um mesmo grupo (nível 2), porém corrigido pela variável W_j.

3.1.3 Modelo de Regressão com Coeficientes Aleatórios

Neste modelo pode-se considerar o intercepto (β_{0j}) e o coeficiente de inclinação (β_{1j}), variando por grupo, ou seja, podem ser considerados como coeficientes aleatórios. Considerando que a variável resposta é Y e uma única variável explanatória do nível 1 é X, então o modelo do nível 1 é da forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{3.7}$$

com $i = 1, 2, \ldots, n_j$ e $j = 1, 2, \ldots, J$ em que

β_{0j} é o intercepto para a j-ésima unidade do nível 2, e representa o valor esperado da variável resposta Y_{ij} quando X_{ij} for igual a zero;

β_{1j} é a inclinação associada a variável explicativa X_{ij} da i-ésima unidade do nível 1 para a j-ésima unidade do nível 2;

e e_{ij} é definido como em (3.1).

Para os modelos de regressão no nível 2, os coeficientes de regressão são considerados como variáveis resposta, temos:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{3.8}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{3.9}$$

em que

γ_{00} é o valor esperado dos interceptos dos J grupos;

γ_{10} é o valor esperado das inclinações dos J grupos;

u_{0j} é o efeito aleatório da j-ésima unidade do nível 2 no intercepto β_{0j};

u_{1j} é o efeito aleatório da j-ésima unidade do nível 2 na inclinação β_{1j};

$u_{0j} \sim N(0, \tau_{00})$ e $u'_{0j}s$ independentes;

$u_{1j} \sim N(0, \tau_{11})$ e $u'_{1j}s$ independentes;

e $u'_{0j}s$ e $u'_{1j}s$ independentes dos $e'_{ij}s$.

A matriz de variâncias e covariâncias dos efeitos aleatórios do nível 2 pode ser escrita como :

$$Var = \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix} = T,$$

em que

$\tau_{00} = Var(u_{0j})$ é a variância não condicional dos interceptos;

$\tau_{11} = Var(u_{1j})$ é a variância não condicional das inclinações;

$\tau_{01} = Cov(u_{0j}, u_{1j})$ é a covariância não condicional entre interceptos e inclinações;

Os componentes de variância e covariância são chamados de não condicionais, uma vez que o modelo não apresenta preditor no nível 2.

Quando substituímos as equações (3.8) e (3.9) na equação (3.7), temos o modelo combinado:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}, \tag{3.10}$$

com $i = 1, 2, \ldots, n_j$ e $j = 1, 2, \ldots, J$.

Neste modelo Y_{ij} é composto por $\gamma_{00}+\gamma_{10}X_{ij}$ mais uma parte aleatória com os seguintes componentes:

u_{0j} é o efeito do j-ésimo grupo sobre a média;

$u_{1j}X_{ij}$ onde u_{1j} é o efeito aleatório do j-ésimo grupo sobre a inclinação β_{1j};

e e_{ij} que é o erro aleatório do nível 1.

3.1.4 Interceptos e Inclinações como Respostas

Para este tipo de modelo incorporamos variáveis (W_j) no modelo do nível 2 de forma que elas ajudem a explicar não só a variabilidade dos interceptos, mas também a das inclinações. Desta forma as equações (3.8) e (3.9) serão substituídas por:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \tag{3.11}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, \tag{3.12}$$

em que

γ_{00} é o valor esperado do intercepto para W_j igual a zero;

γ_{01} é o coeficiente de regressão associado a variável explicativa W_j do nível 2 relativo ao intercepto;

γ_{11} é o coeficiente de regressão associado a variável explicativa W_j do nível 2 relativo à inclinação;

u_{0j} é o efeito aleatório da j-ésima unidade do nível 2 sobre o intercepto para W_j igual a zero;

u_{1j} é o efeito aleatório da j-ésima unidade do nível 2 sobre a inclinação para W_j igual a zero;

$u_{0j} \sim N(0, \tau_{00})$ e $u'_{0j}s$ independentes;

$u_{1j} \sim N(0, \tau_{11})$ e $u'_{1j}s$ independentes;

e $u'_{0j}s$ e $u'_{1j}s$ independentes dos $e'_{ij}s$;

$\tau_{00} = Var(u_{0j})$ é a variância populacional dos interceptos corrigida pela variável W_j;

$\tau_{11} = Var(u_{1j})$ é a variância populacional das inclinações corrigida pela variável W_j;

$\tau_{01} = Cov(u_{0j}, u_{1j})$ é a covariância entre β_{0j} e β_{1j}.

τ_{00}, τ_{11} e τ_{01} são agora componentes de variância e covariância condicionais ou residuais em β_{0j} e β_{1j} depois de incluída W_j.

Substituindo as equações (3.11) e (3.12) na equação (3.7) , tem-se:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij} + e_{ij}, \tag{3.13}$$

com $i = 1, 2, \ldots, n_j$ e $j = 1, 2, \ldots, J$.

O modelo combinado (3.13) envolve as variáveis explicativas X_{ij} do nível 1 e W_j do nível 2, sendo $\gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}W_j+\gamma_{11}W_jX_{ij}$ a parte fixa ou determinística do modelo e o segmento $u_{0j} + u_{1j}X_{ij} + e_{ij}$ contém todos os termos aleatórios do modelo, sendo chamado de parte aleatória ou estocástica do modelo.

As variáveis explicativas X e W dos níveis 1 e 2, respectivamente também podem ser consideradas centradas na média amostral global. Centrar as variáveis explicativas na média amostral global pode ser adequado para a interpretação do intercepto de regressão β_{0j}, quando, por exemplo, o valor zero não for adequado para as variáveis explicativas do nível 1 incluídas no modelo.

Alguns submodelos são decorrentes de mudanças na equação (3.12) que são:

- **ANCOVA com 1 Fator e Efeitos Aleatórios.** Este modelo é obtido quando considera-se que as inclinações não variam aleatoriamente e não são afetadas pelo efeito de W_j, que é uma característica do grupo. A equação (3.12) torna-se:

  $$\beta_{1j} = \gamma_{10},$$

  com $j = 1, \ldots, J$.

- **Modelo com Inclinações Variando Não Aleatoriamente.** Obtemos este modelo quando a variância residual (τ_{11}) é bem próxima de zero. A equação (3.12) é dada por:

  $$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j,$$

  com $j = 1, \ldots, J$.

3.1.5 Forma Geral do Modelo

A extensão para modelos com múltiplos preditores em ambos os níveis é bastante simples. As expressões gerais para modelos lineares hierárquicos com 2 níveis, considerando que existem q variáveis explicativas no nível 1 ($q = 1, \ldots, Q$) e p variáveis explicativas no nível 2 ($p = 1, \ldots, P$) são dadas por:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{Qj}X_{Qij} + e_{ij}, \tag{3.14}$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_{1j} + \gamma_{q2}W_{2j} + \ldots + \gamma_{qP}W_{Pj} + u_{qj} = \gamma_{q0} + \sum_{p=1}^{P} \gamma_{qp}W_{pj} + u_{qj}, \tag{3.15}$$

com $i = 1, 2 \ldots, n_j$, *j=1,2,...,J, q=0,1,...,Q e p=1,...,P.*

A equação (3.14) é correspondente ao nível 1 e a equação (3.15) é correspondente ao nível 2.

É importante salientar que a inclusão de variáveis explicativas nas equações do modelo do nível 2, com exceção da que representa o coeficiente β_{0j}, resulta no aparecimento de termos de interação entre variáveis dos dois níveis do modelo.

3.2 Métodos de Estimação dos Parâmetros

Segundo Bryk & Raudenbush(2002) existem três tipos de parâmetros que podem ser estimados em um modelo linear hierárquico com 2 níveis, são eles: efeitos fixos, coeficientes aleatórios do nível 1 e componentes de variância e covariância.

Considerando agora o modelo mais geral obtido pelas equações (3.14) e (3.15), a extensão dos princípios básicos de estimação é feita de forma direta. O modelo geral do nível 1 com Q variáveis pode ser escrito em notação matricial da forma

$$\underset{\sim}{Y}_j = \mathbf{X}_j\underset{\sim}{\beta}_j + \underset{\sim}{e}_j, \quad j = 1, \ldots, J, \tag{3.16}$$

sendo

$$\underset{\sim}{Y}_j' = \left[\begin{array}{cccc} Y_{1j} & Y_{2j} & \ldots & Y_{n_j j} \end{array} \right],$$

o vetor da variável resposta do grupo j;

$$\mathbf{X}_j = \left[\begin{array}{cccc} 1 & X_{11j} & \ldots & X_{Q1j} \\ 1 & X_{12j} & \ldots & X_{Q2j} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n_j j} & \ldots & X_{Qn_j j} \end{array} \right],$$

a matriz de variáveis preditoras do nível 1, do grupo j;

$$\underset{\sim}{\beta}_j' = \left[\begin{array}{cccc} \beta_{0j} & \beta_{1j} & \ldots & \beta_{Qj} \end{array} \right],$$

o vetor de parâmetros desconhecidos e,

$$\underset{\sim}{e}_j' = \left[\begin{array}{cccc} e_{1j} & e_{2j} & \ldots & e_{n_j j} \end{array} \right],$$

o vetor de erros aleatórios.

O modelo geral para $\underset{\sim}{\beta}_j$ no nível 2 é:

$$\underset{\sim}{\beta}_j = \mathbf{W}_j \underset{\sim}{\gamma} + \underset{\sim}{u}_j, \quad j = 1, \ldots, J, \tag{3.17}$$

em que

$$\mathbf{W}_j = \left[\begin{array}{cccc} \underset{\sim}{W}_{0j} & \underset{\sim}{0} & \ldots & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{W}_{1j} & \ldots & \underset{\sim}{0} \\ \vdots & \vdots & \ldots & \vdots \\ \underset{\sim}{0} & \underset{\sim}{0} & \ldots & \underset{\sim}{W}_{Qj} \end{array} \right],$$

é a matriz de variáveis preditoras do nível 2, sendo $\underset{\sim}{W}_{qj} = \left[\begin{array}{cccc} 1 & W_{1j} & W_{2j} & \ldots & W_{Pj} \end{array} \right]$ o vetor de preditores de β_{qj}, e $\underset{\sim}{0}$ um vetor $P \times 1$ de zeros, $q = 0, 1, \ldots, Q,\ p = 1, \ldots, P$ e $j = 1, \ldots, J$;

$$\underset{\sim}{\gamma}' = \left[\begin{array}{cccc} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_Q \end{array} \right],$$

é o vetor de efeitos fixos, sendo $\underset{\sim}{\gamma}_q = \left[\begin{array}{cccc} \gamma_{q0} & \gamma_{q1} & \gamma_{q2} & \ldots & \gamma_{qP} \end{array} \right]$ e ;

$$\underset{\sim}{u}'_j = \left[\begin{array}{cccc} u_{0j} & u_{1j} & \ldots & u_{Qj} \end{array} \right],$$

é o vetor de efeitos aleatórios.

Dessa forma combinando as equações (3.16) e (3.17) temos o modelo combinado

$$\underset{\sim}{Y}_j = \mathbf{X}_j \mathbf{W}_j \underset{\sim}{\gamma} + \mathbf{X}_j \underset{\sim}{u}_j + \underset{\sim}{e}_j, \quad j = 1, \ldots, J. \tag{3.18}$$

Considerando $\mathbf{A}_j = \mathbf{X}_j \mathbf{W}_j$, o modelo pode ser escrito na forma

$$\underset{\sim}{Y}_j = \mathbf{A}_j \underset{\sim}{\gamma} + \mathbf{X}_j \underset{\sim}{u}_j + \underset{\sim}{e}_j, \quad j = 1, \ldots, J. \tag{3.19}$$

As suposições para este modelo são:

$$\underset{\sim}{e}_j \sim N(0, \mathbf{R}), \quad \mathbf{R} = \sigma^2 \mathbf{I}_{n_j},$$

onde $\mathbf{I}_{n_j}$ é a matriz identidade de dimensão n_j, $j = 1, \ldots, J$; e $\underset{\sim}{u}_j \sim N(0, \mathbf{G})$,

onde

$$\mathbf{G} = \left[\begin{array}{cccc} \tau_{00} & \tau_{01} & \ldots & \tau_{0Q} \\ \tau_{10} & \tau_{11} & \ldots & \tau_{1Q} \\ \vdots & \vdots & \ldots & \vdots \\ \tau_{Q0} & \tau_{Q1} & \ldots & \tau_{QQ} \end{array} \right],$$

é a matriz de variância e covariância.

Em (3.16), se $\underset{\sim}{Y}$ é um vetor de observações $n \times 1$ com matriz de variância e covariância $\mathbf{V}$, Goldstein(1999) mostra que, se $\mathbf{V}$ é conhecida, então o estimador do parâmetro $\underset{\sim}{\beta}$ é dado por:

$$\underset{\sim}{\hat{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\underset{\sim}{Y}, \quad cov(\underset{\sim}{\hat{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}, \tag{3.20}$$

que são os estimadores de mínimos quadrados generalizados usuais.

Se $\underset{\sim}{\beta}$ é conhecido mas $\mathbf{V}$ é desconhecido, pode-se obter os estimadores $\underset{\sim}{\beta}^*$ dos parâmetros de $\mathbf{V}$ usando novamente o método de mínimos quadrados generalizados como:

$$\underset{\sim}{\hat{\beta}^*} = (\mathbf{X}^{*'}(\mathbf{V}^*)^{-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}(\mathbf{V}^*)^{-1}\underset{\sim}{Y}^*, \tag{3.21}$$

em que $\underset{\sim}{Y}^*$ é o vetor de elementos da matriz triangular superior $(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta})(\underset{\sim}{Y} - \mathbf{X}\underset{\sim}{\beta})'$, que é uma matriz quadrada, $\mathbf{V}^*$ é a matriz de variância e covariância de $\underset{\sim}{Y}^*$ e $\mathbf{X}^*$ é a matriz do delineamento que conduz $\underset{\sim}{Y}^*$ a $\mathbf{V}$ na regressão de $\underset{\sim}{Y}^*$ em $\mathbf{X}^*$.

Quando nem $\underset{\sim}{\beta}$ e nem $\mathbf{V}$ são conhecidos, os estimadores são obtidos de modo que as equações (3.20) e (3.21) são satisfeitas simultaneamente, fornecendo, assim, os estimadores de mínimos quadrados generalizados iterativos. Goldstein(1999) mostra também que os estimadores produzidos por esse método são equivalentes aos estimadores de máxima verossimilhança, quando se supõe a distribuição normal para os dados. O processo de estimação começa com uma estimativa inicial de $\mathbf{V}$ usada para obter $\underset{\sim}{\hat{\beta}}$, e então obter uma nova estimativa melhorada de $\mathbf{V}$, e assim por diante, até que a convergência do processo seja obtida.

Sulivan et al.(1999) apresentam o processo de estimação do vetor de parâmetros fixos $\underset{\sim}{\gamma}$, para o vetor de efeitos aleatórios, para os componentes de variância e covariância $\mathbf{R}$ e $\mathbf{G}$ e para os efeitos aleatórios u.

- *Estimação de efeitos fixos.*

Para o modelo combinado (3.19) a estimação dos efeitos fixos pode ser feita utilizando o método de mínimos quadrados ponderados ou por mínimos quadrados generalizados, dado por:

$$\underset{\sim}{\hat{\gamma}} = (\mathbf{A}'\hat{\mathbf{V}}^{-1}\mathbf{A})^{-1}\mathbf{A}'\hat{\mathbf{V}}^{-1}\underset{\sim}{Y}, \tag{3.22}$$

com

$$\mathbf{V} = var(\underset{\sim}{Y}) = \mathbf{X}\mathbf{G}\mathbf{X}' + \mathbf{R},$$

em que $\mathbf{A}$ é uma matriz $N \times J$ com $N = \sum_{j=1}^{J} n_j$, e $\hat{\mathbf{V}}$ é a estimativa da matriz $\mathbf{V}$ com $\mathbf{G}$ e $\mathbf{R}$ substituídos pelos seus respectivos estimadores de máxima verossimilhança.

A variância dos estimador $\underset{\sim}{\hat{\gamma}}$ é estimada por:

$$\widehat{var}(\hat{\underset{\sim}{\gamma}}) = (\mathbf{A}'\hat{\mathbf{V}}^{-1}\mathbf{A})^{-1}. \tag{3.23}$$

- *Estimação dos componentes de variância e covariância.*

Se os tamanhos das amostras n_j são todos iguais, existem expressões fechadas para estimar os parâmetros de variância e covariância. No entanto se os n_j são diferentes são utilizados métodos numéricos iterativos para obter as estimativas. Normalmente esses métodos são baseados em técnicas de estimação por máxima verossimilhança(MV). As estimativas de máxima verossimilhança de $\mathbf{G}$ e $\mathbf{R}$ são encontradas maximizando a função de log-verossimilhança dada por

$$l_{MV}(GR) = -\frac{1}{2}log\,|\mathbf{V}| - \frac{N}{2}log(\underset{\sim}{r}'\mathbf{V}^{-1}\underset{\sim}{r}) - \frac{N}{2}\left[1 + log\frac{2\pi}{N}\right], \tag{3.24}$$

onde

$$\underset{\sim}{r} = \underset{\sim}{Y} - \mathbf{A}(\mathbf{A}'\mathbf{V}^{-1}\mathbf{A})'\mathbf{A}'\mathbf{V}^{-1}\underset{\sim}{Y}.$$

Se o número J de unidades do nível 2 é grande, então, os estimadores gerados pela máxima verossimilhança são aproximadamente iguais aos gerados pela máxima verossimilhança restrita. Os estimadores de máxima verossimilhança restrita(MVR) para os componentes de variância e covariância são baseados nos resíduos, os quais são obtidos após a estimativa dos efeitos fixos (3.22) através dos métodos de mínimos quadrados ponderados ou mínimos quadrados generalizados. Nota-se que o estimador de MVR leva em conta o número de graus de liberdade usado nas estimativas dos efeitos fixos quando estima-se os componentes de variância e covariância. As estimativas de máxima verossimilhança restrita de $\mathbf{G}$ e $\mathbf{R}$ são encontradas maximizando a seguinte função de log-verossimilhança

$$l_{MVR}(GR) = -\frac{1}{2}log\,|\mathbf{V}| - \frac{1}{2}log\left|\mathbf{A}'\mathbf{V}^{-1}\mathbf{A}\right| - \frac{(N-p)}{2}log(\underset{\sim}{r}'\mathbf{V}^{-1}\underset{\sim}{r}) - \frac{(N-p)}{2}\left[1 + log\frac{2\pi}{(N-p)}\right] \tag{3.25}$$

onde

$$\underset{\sim}{r} = \underset{\sim}{Y} - \mathbf{A}(\mathbf{A}'\mathbf{V}^{-1}\mathbf{A})'\mathbf{A}'\mathbf{V}^{-1}\underset{\sim}{Y}$$

e $p = rank(\mathbf{A})$.

- *Estimação dos efeitos aleatórios.*

As estimativas dos efeitos aleatórios podem ser obtidas substituindo (3.22) na equação obtida quando derivamos l_{MV} em relação a $\underset{\sim}{\gamma}$ e $\underset{\sim}{u}_j$. Dessa forma temos que :

$$\hat{\underset{\sim}{u}}_j = \mathbf{G}\mathbf{X}'\hat{\mathbf{V}}^{-1}(\underset{\sim}{Y} - \mathbf{A}\hat{\underset{\sim}{\gamma}}). \tag{3.26}$$

3.3 Testes de Hipóteses

Nesta seção abordaremos os testes de hipóteses para os efeitos fixos, efeitos aleatórios e para os componentes de variância e covariância de um Modelo Linear Hierárquico.

Apresentaremos testes para hipóteses que envolvem um único parâmetro e testes para múltiplos parâmetros.

3.3.1 Teste de Hipóteses para Efeitos Fixos

É interessante investigar em um modelo linear hierárquico se cada um de seus efeitos fixos estimados são significativamente diferentes de zero. A hipótese de interesse para testar um único efeito fixo é:

$$H_0 : \gamma_{qp} = 0.$$

A estatística de teste é calculada considerando o estimador de máxima verossimilhança ou a máxima verossimilhança restrita para o erro padrão, dada por:

$$t = \frac{\widehat{\gamma}_{qp}}{\sqrt{\widehat{V}ar(\widehat{\gamma}_{qp})}} \tag{3.27}$$

onde $\widehat{\gamma}_{qp}$ é o estimador do efeito fixo γ_{qp} e $\widehat{V}ar(\widehat{\gamma}_{qp})$ é a estimativa da variância de $\widehat{\gamma}_{qp}$.

Esta estatística segue uma distribuição t-Student com $J - P - 1$ graus de liberdade para dados balanceados e para algumas situações de dados não balanceados. Na maioria das situações, a distribuição de (3.27) é aproximada, bem como os graus de liberdade.

Para testar mais de um efeito fixo simultaneamente, é necessário fazer o uso de contrastes e testá-los através da estatística de Wald.

Suponha que o vetor de efeitos fixos seja dado por:

$$\underset{\sim}{\gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{21} & \gamma_{22} \end{pmatrix}'.$$

Então, por exemplo, se quisermos testar a hipótese

$$\begin{aligned} H_0 : \gamma_{12} &= 0 \\ \gamma_{22} &= 0, \end{aligned}$$

é conveniente reescrevê-la na forma matricial

$$H_0 : \mathbf{C}'\underset{\sim}{\gamma} = 0$$

onde

$$\mathbf{C}' = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Considerando $\hat{\underset{\sim}{\gamma}}$ o estimador do vetor de efeitos fixos $\underset{\sim}{\gamma}$ e $\widehat{\mathbf{V}}_{\underset{\sim}{\gamma}}$ a estimativa da variância de $\hat{\underset{\sim}{\gamma}}$. Logo

$$\widehat{\mathbf{V}}ar(\mathbf{C}'\hat{\underset{\sim}{\gamma}}) = \mathbf{C}'\widehat{\mathbf{V}}_{\hat{\underset{\sim}{\gamma}}}\mathbf{C} = \widehat{\mathbf{V}}_{\mathbf{C}}$$

e sob H_0 a estatística do teste de Wald é dada por:

$$H = \hat{\underset{\sim}{\gamma}}'\mathbf{C}\widehat{\mathbf{V}}_{\mathbf{C}}^{-1}\mathbf{C}'\hat{\underset{\sim}{\gamma}}$$

cuja distribuição assintótica é uma distribuição χ^2 com número de graus de liberdade igual ao número de linhas da matriz $\mathbf{C}$.

O teste da razão de verossimilhança é uma outra abordagem para o teste multiparamétrico que pode ser usado quando se trabalha com o método de máxima verossimilhança completo, não sendo aplicável para o método de máxima verossimilhança restrito (Bryk & Raudenbush, 2002).

3.3.2 Teste de Hipóteses para Efeitos Aleatórios

A hipótese de interesse é da forma:

$$H_0 : \beta_{qj} = 0,$$

ou equivalentemente

$$H_0 : u_{qj} = 0.$$

A estatística de teste é obtida fazendo a razão entre a estimativa do efeito aleatório $\widehat{u}_{qj}$ pela estimativa do seu erro padrão, da seguinte forma:

$$t = \frac{\widehat{u}_{qj}}{\sqrt{\widehat{V}ar(\widehat{u}_{qj})}} \tag{3.28}$$

A estatística (3.28) tem distribuição t-Student para dados balanceados e para algumas situações de dados não balanceados. A estimativa do erro padrão ($\sqrt{\widehat{V}ar(\widehat{u}_{qj})}$) é maior quando se usa o o método de máxima verossimilhança restrito de que quando se usa o método de máxima verossimilhança completo, especialmente quando o número de unidades do nível 2, J , é pequeno (Sulivan *et al*, 1999).

No caso multiparamétrico, vamos considerar $\boldsymbol{\beta}$ como sendo o vetor de parâmetros aleatórios com dimensão $J(Q+1) \times 1$, logo a hipótese linear geral associada a $\boldsymbol{\beta}$ é:

$$H_0 : \mathbf{C}'\boldsymbol{\beta} = 0. \tag{3.29}$$

Se o vetor de parâmetros pode ser estimado por mínimos quadrados ordinários(MQO), então a hipótese linear geral (3.29) pode ser testada através da estatística:

$$H_{MQO} = \hat{\boldsymbol{\beta}}' \mathbf{C} (\mathbf{C}' \widehat{\mathbf{V}} \mathbf{C})^{-1} \mathbf{C}' \hat{\boldsymbol{\beta}}', \tag{3.30}$$

onde $\widehat{\mathbf{V}}$ é uma matriz bloco diagonal com cada bloco de dimensão $(Q+1) \times (Q+1)$ e igual a :

$$\widehat{\mathbf{V}}_j = \hat{\sigma}^2 (\mathbf{X}_j' \mathbf{X}_j)^{-1}.$$

3.3.3 Teste de Hipóteses para Componentes de Variância e Covariância

Assim como os efeitos fixos e aleatórios, os componentes de variância podem também ser testados isoladamente ou simultaneamente.

Seja τ_{qq} a variância dos coeficientes u_{qj}. Para avaliar se este coeficiente deve ser considerado fixo ou aleatório, podemos testar a hipótese nula:

$$H_0 : \tau_{qq} = 0 \tag{3.31}$$

Para testar esta hipótese podemos utilizar a estatística:

$$z = \frac{\widehat{\tau}_{qq}}{\sqrt{\widehat{V}ar(\widehat{\tau}_{qq})}}. \tag{3.32}$$

A estatística (3.32) tem distribuição aproximadamente normal, para grandes amostras. A estimativa do erro padrão de $\widehat{\tau}_{qq}$ é calculada através do inverso da matriz de informação de Fisher.

Uma outra possibilidade para testar a hipótese (3.31) é utilizar as estimativas de mínimos quadrados ordinários (MQO). Este teste só é possível se todos ou pelo menos a maioria dos grupos tiver dados suficientes para calcular as estimativas de MQO. Dessa forma, considerando o modelo

$$\beta_{qj} = \gamma_{q0} + \sum_{p=1}^{P} \gamma_{qp} W_{pj}, \tag{3.33}$$

este teste pode ser ser realizado através da estatística

$$\sum_{j=1}^{J} \frac{(\widehat{\beta}_{qj} - \widehat{\gamma}_{q0} - \sum_{p=1}^{P} \widehat{\gamma}_{qp} W_{pj})^2}{\widehat{V}_{qqj}} \tag{3.34}$$

cuja distribuição é aproximadamente χ^2 com $J - P - 1$ graus de liberdade.

Para testar um ou mais componentes de variância e covariância simultaneamente, pode-se utilizar o teste de razão de verossimilhança, onde a hipótese nula a ser testada é:

$$H_0 : \mathbf{T} = \mathbf{T}_0 \tag{3.35}$$

contra a alternativa

$$H_1 : \mathbf{T} = \mathbf{T}_1, \tag{3.36}$$

onde $\mathbf{T}_0$ é uma forma reduzida de $\mathbf{T}_1$.

Para calcular a estatística do teste da razão de verossimilhança é preciso obter a "Deviance" do modelo ajustado sob $H_0(D_0)$ e a "Deviance" do modelo ajustado sob $H_1(D_1)$.

Esta "Deviance" é a mesma definida na seção 2.2. A "Deviance" pode ser vista como uma medida de qualidade do ajuste. A estatística do teste da razão de verossimilhança é dada por:

$$H = D_0 - D_1$$

e tem distribuição χ^2 com m graus de liberdade, onde m corresponde à diferença do número de componentes de variância e covariância estimados nos dois modelos sob as hipóteses H_0 e H_1.

CAPÍTULO 4

Modelos Lineares Generalizados Hierárquicos

Neste capítulo abordaremos os Modelos Lineares Generalizados Hierárquicos (MLGH), também conhecidos como modelos lineares generalizados mistos segundo Breslow & Clayton(1993) ou modelos lineares generalizados com efeitos aleatórios (Schall, 1991). No capítulo 3 mostramos que os modelos lineares hierárquicos prestam-se bem aos casos em que se supõem variáveis resposta normalmente distribuídas em cada um dos níveis. Entretanto em muitas aplicações de regressão a variável resposta de interesse é qualitativa, com dois ou mais resultados possíveis, ou uma variável de contagem, muito comum em certos tipos de estudos, como os epidemiológicos e, da mesma forma deseja-se estimar essa resposta não só em termos de características individuais, mas também de grupos. Essa generalização em termos de aplicação pode ser alcançada pelos MLGH.

Empregaremos neste capítulo os conceitos já estudados dos MLG e MLH, de forma que seja alcançada e apresentada a generalização hierárquica segundo o enfoque de Bryk & Raudenbush(2002).

Observa-se que os MLGH apresentam a mesma estruturação do MLH. O nível 1 do MLGH consiste em três partes distintas, a saber: um modelo amostral, uma função de ligação e um modelo estrutural. Assim como nos MLG a variáveis resposta dos MLGH

podem seguir as diferentes distribuições da família exponencial, a depender do interesse do estudo.

Neste trabalho a variável resposta de interesse segue uma distribuição Poisson. Portanto o caso específico do modelo Poisson será apresentado na seção 4.2.

Para uma melhor percepção da relação entre MLH e MLGH apresentamos na Seção 4.1 o caso do MLH como um caso particular do MLGH.

4.1 O Modelo Hierárquico Normal

O nível 1 do MLGH consiste em três partes: Um modelo amostral, uma função de ligação e um modelo estrutural. De fato o MLH pode ser escrito como um caso especial do MLGH, onde o modelo amostral é normal, a função de ligação é a função identidade e o modelo estrutural é linear.

1. Nível 1 (*Modelo Amostral*)

   O modelo amostral para o MLH com dois níveis pode ser escrito como

   $$Y_{ij} \mid \mu_{ij} \sim NID(\mu_{ij}, \sigma^2), \tag{4.1}$$

   ou seja, a variável resposta Y_{ij} do nível 1, dado o valor predito μ_{ij} é normal e independentemente distribuída com um valor esperado μ_{ij} e uma variância constante, σ^2.

2. (*Função de Ligação*)

   Em geral, é possível transformar o valor predito μ_{ij} do nível 1, para garantir que as predições estejam condicionadas a variar dentro de um determinado intervalo. Denotaremos esta transformação por η_{ij} e chamaremos de função de ligação. No caso da Normal esta transformação não é necessária, entretanto podemos escrever esta não transformação como

$$\eta_{ij} = \mu_{ij} \tag{4.2}$$

que é conhecida como "função identidade".

3. (*Modelo Estrutural*)

   A transformação do valor predito η_{ij} agora está relacionado com os preditores do modelo linear por meio do modelo estrutural

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots + \beta_{Qj}X_{Qij}. \tag{4.3}$$

   Combinando (4.1), (4.2) e (4.3) temos o nível 1 do MLH.

4.2 O Modelo Hierárquico Poisson

1. Nível 1 (*Modelo Amostral*)

   Seja Y_{ij} o número de eventos ocorridos durante um intervalo de tempo m_{ij} que pode ser denominado de *taxa de exposição.* Por exemplo Y_{ij} pode ser o número de crimes que uma pessoa i numa vizinhança j cometeu durante 5 anos, então $m_{ij} = 5$. Escrevemos

$$Y_{ij} \mid \lambda_{ij} \sim P(m_{ij}, \lambda_{ij}) \tag{4.4}$$

   Para denotar que Y_{ij} tem uma distribuição Poisson com taxa de exposição m_{ij} e taxa de evento por período de tempo λ_{ij}. De acordo com a distribuição Poisson, o valor esperado e a variância de Y_{ij}, dado a taxa de evento, λ_{ij}, é dado por

$$E(Y_{ij} \mid \lambda_{ij}) = m_{ij}\lambda_{ij} \qquad Var(Y_{ij} \mid \lambda_{ij}) = m_{ij}\lambda_{ij} \tag{4.5}$$

ou seja, o valor esperado do número de eventos Y_{ij} para a unidade i do grupo j é a taxa de evento λ_{ij} multiplicada pela taxa de exposição m_{ij}, e a variância é igual a média.

A taxa de exposição m_{ij} não precisa ser uma medida de tempo. Por exemplo, numa aplicação clássica do modelo Poisson, Y_{ij} é o número de bombas lançadas numa vizinhança i de uma cidade j durante uma guerra, e m_{ij} é a área desta vizinhança. Um caso bastante comum surge quando para cada i e j a taxa de exposição é a mesma. Por exemplo, Y_{ij} é o número de crimes cometidos durante um ano por cada pessoa i dentro de cada vizinhança j. Neste caso, $m_{ij} = 1$ e o valor predito de Y_{ij} quando $m_{ij} = 1$ será a taxa de evento λ_{ij}.

2. (*Função de Ligação*)

   A função de ligação canônica quando o modelo amostral do nível 1 é Poisson é a ligação logarítmica, ou seja,

   $$\eta_{ij} = log(\lambda_{ij}). \tag{4.6}$$

   Então, η_{ij} é o log da taxa de evento. Assim, se a taxa de evento for 1 o logaritmo é 0. Quando a taxa de evento é menor que 1 o logaritmo é negativo e quando a taxa de evento é maior que 1 o logaritmo é positivo . Dessa forma, quando λ_{ij} for positivo, $log(\lambda_{ij})$ será um valor real.

3. (*Modelo Estrutural*)

   O modelo estrutural é exatamente igual ao apresentado na equação (4.3). Note que as estimativas dos $\beta's$ na equação (4.3) torna possível a predição do logaritmo da taxa de evento $\hat{\eta}_{ij}$ para qualquer caso. Tal predição pode ser convertida para uma taxa de evento $\hat{\lambda}_{ij}$, calculando $\hat{\lambda}_{ij} = exp\{\hat{\eta}_{ij}\}$. Qualquer que seja o valor de $\hat{\eta}_{ij}$, $\hat{\lambda}_{ij}$ será positivo.

4. Nível 2 (*Modelo*)

   No caso do nível 2, o modelo será o mesmo do caso normal apresentado na equação (3.15), ou seja:

$$\beta_{qj} = \gamma_{q0} + \sum_{p=1}^{P} \gamma_{qp} W_{pj} + u_{qj}, \tag{4.7}$$

   com $i = 1, 2 \ldots, n_j$, *j=1,2,...,J, q=0,1,...,Q e p=1,...,P.*

4.3 Estimação em MLGH

Obter estimadores de máxima verossimilhança para modelos hierárquicos é um problema que consiste em dois passos. O primeiro passo é encontrar a verossimilhança. Este requer a integração da distribuição conjunta dos dados e efeitos aleatórios com respeito aos efeitos aleatórios. O segundo passo é a maximização da verossimilhança. O primeiro passo é fácil quando o modelo é linear com efeitos aleatórios normalmente distribuídos em cada nível. O único problema é a maximização. Contudo, em MLGH esta situação se complica, pois eles são formados por um modelo amostral de nível 1 não Normal e por um modelo estrutural de nível 2 contando com efeitos aleatórios normalmente distribuídos.

Considerando Y como sendo um vetor de dados, u como sendo um vetor de efeitos aleatórios, e ω um vetor de parâmetros. Os parâmetros incluem os componentes de variância e covariância e os coeficientes fixos da regressão. A distribuição dos dados dado os efeitos aleatórios do nível 1 do modelo é denotada por $f(Y/u, \omega)$, enquanto a distribuição dos efeitos aleatórios é denotada por $p(u/\omega)$. A distribuição conjunta dos dados e efeitos aleatórios é dada por:

$$g(Y, u/\omega) = f(Y/u, \omega)p(u/\omega),$$

enquanto a verossimilhança dos dados dado o vetor de parâmetros ω é a densidade marginal de Y, isto é, a integral da distribuição conjunta com relação aos efeitos aleatórios:

$$L(Y/\omega) = \int f(Y/u, \omega)p(u/\omega)du. \tag{4.8}$$

No MLH os dois fatores do produto no integrando são normalmente distribuídos, e isso implica na normalidade da função. Essa situação facilita a dedução da função de verossimilhança e o problema maior passa a ser sua maximização. Já no caso de um MLGH, a questão é que a primeira parcela do integrando é agora Binomial, Poisson ou Multinomial, sendo especificamente Poisson no caso prático abordado nesta pesquisa, acarretando que a distribuição conjunta interna à integral em (4.8) é, na verdade, mista, por exemplo, Poisson- Gamma. A consequência é que a função de verossimilhança em MLGH tem de ser encontrada de forma numérica, assim como sua maximização.

Uma das abordagens para se obter a função de verossimilhança e sua maximização em MLGH é a Penalized Quasi-Likelihood(PQL). Em relação à terminologia "Quasi-Likelihood", o que se pode dizer, de acordo com Zorn(1998) é que: *Enquanto máxima verossimilhança padrão requisita a especificação da distribuição condicional da variável resposta, quasi-likelihood necessita apenas que se defina o relacionamento entre o valor esperado da resposta e as covariáveis, e entre a média e a variância da variável resposta.* Maiores detalhes em Quasi-Likelihood podem ser encontrados em Heyde(1997).

Embora PQL seja o método mais utilizado pela literatura de referência em MLGH, Goldstein(1999), a aproximação do integrando em 4.8 pelo método de Laplace é também uma alternativa viável para realização das inferências em MLGH Bryk & Raudenbush(2002). As estimativas dos parâmetros do modelo apresentado neste trabalho foram obtidas pelo método de aproximação de Laplace.

CAPÍTULO 5

Aplicação

5.1 Introdução

Nos últimos anos tem sido crescente o interesse de especialistas e autoridades governamentais por indicadores quantitativos que, além de auxiliar o entendimento da dinâmica de ciência e tecnologia (C&T) são objeto de estudo de várias áreas do conhecimento, sendo usados tanto para o planejamento e a execução de políticas para o setor como também para que a comunidade científica conheça melhor o sistema no qual está inserida.

Os estudos de produção científica enfrentam desafios. De fato, a produção científica é parte de um grande sistema social que é a ciência. Como afirma Macias-Chapula(1998) "[...] *a ciência necessita ser considerada como um amplo sistema social, no qual uma de suas funções é disseminar conhecimentos. Sua segunda função é assegurar a preservação de padrões e, a terceira, é atribuir crédito e reconhecimento para aqueles cujos trabalhos têm contribuído para o desenvolvimento das idéias em diferentes campos*".

Os indicadores de produção científica são construídos basicamente pela contagem do número de publicações por tipo de documento (livros, artigos, publicações científicas, relatórios, etc.), por instituição, área de conhecimento, país, etc. Tais indicadores procuram

refletir características de produção ou do esforço empreendido, mas não mede a qualidade das publicações. As leis de Lotka, Zipf e Bradford, comentadas no Capítulo 1, também são consideradas como indicadores de produção científica.

Dentre os vários tipos de indicadores bibliométricos que existem, os indicadores de citações estão sendo bastante usados ultimamente. Estes indicadores foram desenvolvidos a partir do princípio de que as referências citadas por um autor identificam de maneira mais precisa o relacionamento entre documentos que tratam do mesmo assunto. Além de serem utilizados por cientistas como instrumentos de recuperação de documentos por assuntos, os indicadores de citações passaram a ser utilizados por sociólogos da ciência e pelos responsáveis pela elaboração de políticas científicas para avaliação da performance dos cientistas(Mugnaini, 2004). Os dois indicadores de citações mais famosos são o Fator de Impacto(FI) e o *índice h*.

Eugene Garfield(1955) propôs o conceito de FI, em seguida Garfield fundou o ISI(Information Sciences Institute), com o objetivo de fornecer informações sobre pesquisas correntes. Desde então o FI e outros índices correlatos têm sido objeto de generalizada controvérsia e de alguns equívocos na sua utilização. O FI de um periódico, em determinado ano, é calculado dividindo-se o número de citações feitas no ano da avaliação (em todos os periódicos indexados) a itens publicados por ele nos dois anos imediatamente anteriores pelo número de artigos (itens-fonte) publicados nesse mesmo periódico também nos dois anos anteriores ao da avaliação. Por exemplo, seja A o número de vezes que artigos publicados em 2005 e 2006 foram citados por revistas indexadas em 2007 e B o número de artigos publicados em 2005 e 2006, o fator de impacto da revista em 2007 será A/B.

O fator de impacto identifica a frequência com que um artigo médio de um periódico é citado em um determinado ano. Pode-se usar este número para avaliar ou comparar a importância relativa de um periódico com outros do mesmo campo ou ver com que

frequência os artigos são citados para determinar quais periódicos são melhores para a sua coleção.

Várias propostas que se baseiam na aritmética entre o número de artigos e o número de citações são conhecidas, mas o método que ganhou fama no meio científico, devido principalmente a sua simplicidade foi o proposto por JE Hirsch(2005). Hirsch propôs um único número, o *índice h*, como uma forma particularmente simples e útil para caracterizar a produção científica de um pesquisador e é definido da seguinte forma: Um cientista tem *índice h* se h de seus artigos publicados(N_p) ao longo de n anos têm pelo menos h citações cada e os $(N_p - h)$ artigos restantes tem uma quantidade de citações menor ou igual a h. Por exemplo um pesquisador que tem índice $h = 115$, significa que ele tem 115 artigos com pelo menos 115 citações cada.

O *índice h* possui limitações e ineficiências, segue algumas destas limitações e ineficiências.

- O *índice h* é limitado pelo número total de publicações. Isto significa que os cientistas com uma curta carreira estão inerentes a uma desvantagem, independentemente da importância das suas descobertas;

- Caso o pesquisador publique muitos reviews, poderá ter h maior que outro que contribua com resultados originais;

- O *índice h* subestima publicações de altíssimo impacto em favor da sustentabilidade da produção científica;

- Embora o *índice h* enfatize o êxito das publicações em favor da produtividade sustentada, ele poderá fazê-lo de forma demasiada. Dois cientistas podem ter o mesmo *índice h*, digamos h = 30, mas um tem 20 artigos que foram citados mais de 1.000 vezes e o outro não tem nenhum artigo com essa quantidade de citações. Claramente a produção científica do primeiro é mais "valiosa".

- Não compara pesquisadores de áreas diferentes;

- É um índice de "passado", não tendo como detectar pesquisadores jovens e brilhantes.

Várias variantes do *índice h* foram propostas para corrigir essas deficiências, mas nenhuma ganhou apoio universal.

Dadas as peculiaridades da ciência, a comunidade científica de cada área ou subárea adota diferentes processos de utilização de veículos de disseminação da produção. Por exemplo, as áreas das ciências exatas e biológicas não têm a mesma cultura de publicação daquelas das ciências sociais. Enquanto as primeiras tendem a privilegiar a publicação de artigos científicos, nas ciências humanas e sociais, privilegia-se a publicação de livros. Esse fato reforça a idéia de que é inadequada a universalização do critério de avaliação da produção científica baseada tão somente em artigos publicados em periódicos especializados.

5.2 Índice de Produção Científica

Neste trabalho foi criado um índice de produção que envolve diversos indicadores de produção científica que são utilizados como veículos de disseminação de produção em diferentes áreas do conhecimento. O objetivo da criação deste índice é ter um instrumento que avalie através de um número a produção científica de todos os professores da UFPE independente da área.

O índice de produção proposto é uma soma ponderada de indicadores de produção. Definir pesos com o compromisso de refletir a percepção e a importância relativa de cada indicador a cada unidade é uma tarefa difícil e questionável. Assim, a obtenção dos pesos atribuídos aos indicadores seguem a portaria normativa Nº 22, de 21 de outubro de 2002 da UFPE, que estabelecia critérios de avaliação do desempenho docente para fim de atribuição da antiga Gratificação de Estimulo à Docência(GED). A Tabela 5.1 mostra

os pesos dos indicadores que compõem o índice. Todos os indicadores que compõem o índice foram obtidos através da soma da produção nos anos de 2004, 2005 e 2006. Tais indicadores foram extraídos do Currículo Lattes de cada professor que estavam atualizados até março de 2008. O Currículo Lattes está disponível na plataforma Lattes (CNPq) http://lattes.cnpq.br. Dado seu grau de abrangência, as informações que constam na Plataforma Lattes são utilizadas tanto no apoio a atividades de gestão, como no apoio à formulação de políticas para a área de ciência e tecnologia. O Currículo Lattes registra a vida passada e atual dos pesquisadores sendo elemento indispensável à análise de mérito e competência dos pleitos pessoais, coletivos ou institucionais, para auxílios financeiros apresentados ao CNPq. A hipótese assumida no presente trabalho é que todos doutores em atividade, estão com seus currículos na Plataforma Lattes atualizados.

Foram considerados 17 indicadores para construir o índice de produção (Tabela 5.1). O primeiro indicador que compõe o índice de produção é o número de orientações concluídas de teses de doutorado, em seguida temos o número de orientações concluídas de dissertações de mestrado. Estes indicadores medem a quantidade de teses e dissertações que o professor orientou entre 2004 e 2006. Também foi considerado o número de livros publicados e o número de capítulos de livros publicados que são outros dois indicadores de produção muito comuns em certas áreas do conhecimento. Outro indicador que compõe o índice de produção é o número total de artigos publicados em periódicos, que é um dos indicadores de produção mais utilizados em todas as áreas do conhecimento. Estes artigos foram classificados em quatro tipos de acordo com o "nível Qualis" de qualidade do periódico. Esta classificação pode ser A, B, C ou sem Qualis. O Qualis foi elaborado pela Capes e consiste, basicamente, na classificação dos veículos de divulgação da produção científica, técnica, artística dos programas de pós-graduação. Mais informações sobre o "Qualis" podem ser encontradas em http://www.capes.gov.br/avaliacao/qualis.

Indicador		Peso
Número de orientações de doutorado		20
Número de orientações de mestrado		15
Número de livros publicados		30
Número de capítulos de livros publicados		15
Número de artigos publicados em periódicos indexados com:	Qualis nível A	30
	Qualis nível B	25
	Qualis nível C	20
	Sem Qualis	15
Número de trabalhos em congressos do tipo:	completo com âmbito de circulação internacional	15
	completo com âmbito de circulação nacional	10
	completo com âmbito de circulação regional	5
	resumo com âmbito de circulação internacional	12
	resumo com âmbito de circulação nacional	8
	resumo com âmbito de circulação regional	3
	resumo expandido com âmbito de circulação internacional	13
	resumo expandido com âmbito de circulação nacional	9
	resumo expandido com âmbito de circulação regional	4

Tabela 5.1: Pesos utilizados para a construção do índice de produção

Por fim também utilizamos o indicador número de trabalhos apresentados em congressos. Classificamos os trabalhos apresentados em congressos em: completos, resumos e resumos expandidos e subdividimos cada um desses a partir do âmbito do evento: Internacional, Nacional e Regional.

Os créditos das produções científicas que foram publicados com mais de um autor foram divididos igualmente entre os autores, ou seja, se um artigo foi publicado com 3 autores cada um deles recebeu 1/3 dos créditos desse artigo. Estabelecido esse critério alguns índices de produção no período dos 3 anos foram números decimais, então aproximamos para o maior inteiro mais próximo quando a parte decimal foi $\geq 0,5$ e para o menor inteiro mais próximo no caso contrário.

Inicialmente foi calculado o índice de produção para todos os professores da UFPE em atividade em dezembro de 2006 num total de 1659 professores entre doutores, mestres, especialistas e graduados em regime de dedicação exclusiva ou não. O índice de produção variou de 0 a 2087, no entanto, a quantidade de zeros foi de 27,5%. Então, para uma melhor modelagem e baseado no fato de que a exigência quanto a produção cientifica é maior entre os professores que possuem doutorado em regime de dedicação exclusiva(DE), foi considerado a sub-amostra de 977 professores que são doutores em regime de dedicação exclusiva. Nesta sub-amostra a quantidade de índices de produção com valor zero foi de 7,3%.

A Figura 5.1 ilustra o histograma dos valores que o índice assume para todos os doutores com DE. A quantidade de professores com índice de produção inferior a 500 corresponde a 90% do total de professores. Menos de 1% dos professores possui índice superior a 1000. A estrutura administrativa da UFPE está subdividida em dez centros acadêmicos e, em cada centro acadêmico existem os departamentos onde se encontram lotados os professores. No CCJ-Centro de Ciências Jurídicas, só identificamos 9 docentes na sub-amostra de interesse e, considerando a necessidade de uma quantidade mínima de elementos amostrais em cada centro, tendo em vista a modelagem estatística a ser adotada,

os docentes do CCJ foram agregados aos docentes do CFCH visto ser este centro o que apresenta valor médio do índice de produção mais próximo do valor médio do índice de produção do CCJ. Na Tabela 5.2 apresentamos a média, o desvio padrão, o menor e o maior índice de produção e o coeficiente de variação(CV) de cada centro. Analisando os valores dos índice de produção por cada um dos 9 centros acadêmicos considerados percebemos que o centro com a maior média do índice de produção é o CIN com uma média de 299,13 e o de menor média é o CE com uma média de 156,42. O CCEN é o segundo com maior média 261,27. O índice estudado apresenta uma dispersão relativa medida pelo CV de 101,19%, destacando-se o CCS com a maior dispersão relativa com CV de 117,94%.

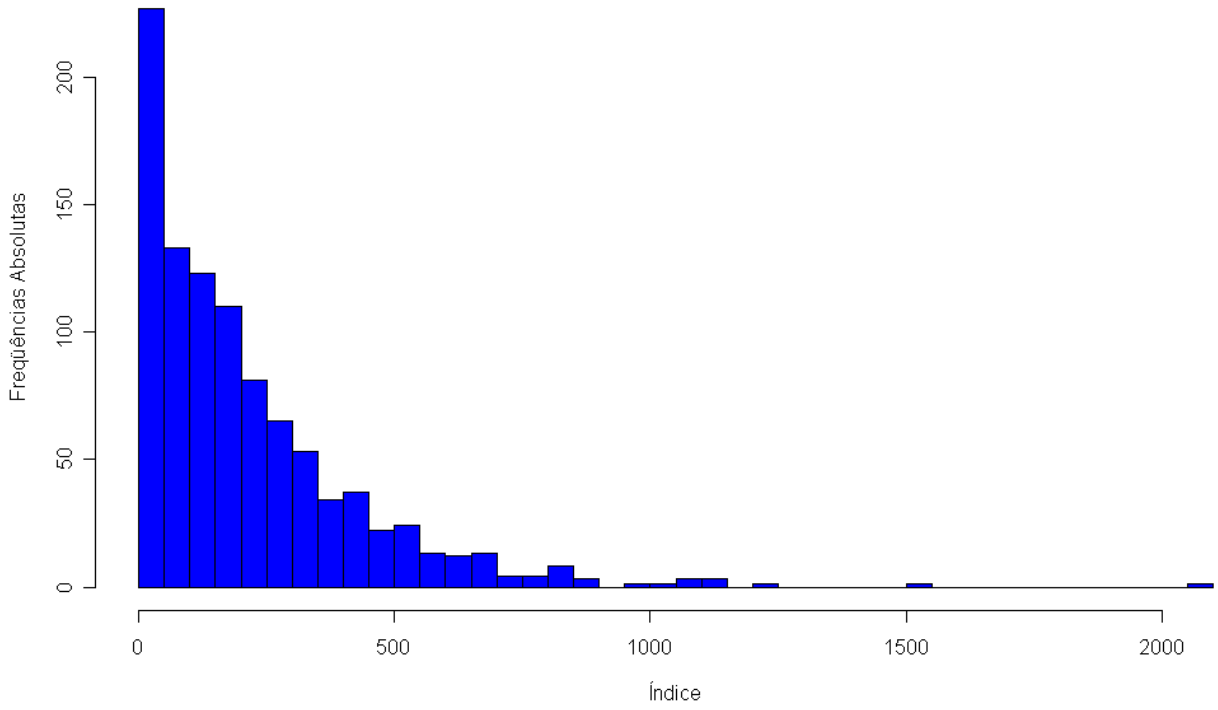

Figura 5.1: Histograma do Índice de produção

CENTRO	Número de Docentes	Mínimo	Máximo	Média	Desvio Padrão(DP)	Coeficiente de Variação(CV)%
CAC	106	0	717	159,67	149,26	93,48
CCB	158	0	1058	204,57	202,67	99,07
CCEN	99	0	2087	261,27	288,81	110,54
CCS	122	0	1217	197,89	233,40	117,94
CCSA	80	0	833	224,56	209,91	93,47
CE	55	0	548	156,42	124,69	79,71
CFCH	109	0	696	192,26	168,64	87,71
CIN	48	0	1036	299,13	244,17	81,62
CTG	200	0	1509	225,73	219,91	97,42
Total	977	0	2087	211,14	213,67	101,19

Tabela 5.2: Estatística descritiva do índice de produção por centro

5.3 Modelagem Através do MLGH

Os dados cadastrais utilizados na modelagem deste trabalho foram cedidos gentilmente pela PROPLAN(Pró-reitoria de Planejamento da UFPE). O banco de dados utilizado na modelagem é formado por 977 professores doutores com dedicação exclusiva com as seguintes variáveis: idade, tempo de serviço, gênero, se é ou não bolsista do CNPq, departamento e centro em que trabalham os professores.

O índice de produção ora descrito é a variável dependente na modelagem descrita a seguir.

5.3.1 Variáveis Explicativas

As variáveis explicativas, consideradas como candidatas a explicar o índice de produção científica no 1° nível hierárquico foram: idade, gênero e tempo de serviço.

Fox(1983) apresenta várias evidências que apontam para uma correlação positiva entre idade e produção científica de pesquisadores. A relação entre idade e produção científica tem gerado muita discussão. Alguns pesquisadores argumentam que a produção científica começa a diminuir muito cedo (Lehman, 1953). Outros acreditam que há uma relação curvilínea e que a produtividade cresce lentamente nas primeiras décadas de vida,

alcançando seu pico na idade intermediária e decresce lentamente nas últimas décadas (Cole, 1987). Por último, outros argumentam que a idade desempenha um papel sem muita importância na produtividade acadêmica, sendo a taxa prévia de publicações o que determina a produtividade posterior. Merton(1968) identificou que à medida que os cientistas envelhecem, menos tempo despendem em pesquisa e que uma grande proporção de seu tempo é gasta em posições administrativas. Contrariando este ponto de vista, Fox(1983) notou que embora o cientista mantenha-se afastado do contexto da pesquisa quando assume cargos administrativos, o aumento de recursos disponíveis facilita, ao invés de impedir, a produtividade científica. Uma das razões apontadas pela literatura para o declínio da produção científica é a de que à medida que os cientistas envelhecem, seu capital financeiro também aumenta, e eles passam a aplicar seu tempo em atividades não relacionadas às atividades de pesquisa.

A partir da variável idade foi criada uma variável dummy que chamamos de **cod_idade**, que assume 1 se a idade for menor ou igual a 50 e 0 caso contrário. O limite 50 foi escolhido porque muitos pesquisadores, tal como Cole(1987) acreditam que a partir desta idade a produção cientifica começa a diminuir. Então o objetivo da escolha dessa variável é verificar se esta queda de produção ocorre com os professores em estudo.

O tempo de serviço (**T_Serv**) na UFPE apresenta uma certa correlação com a idade mas, mesmo assim consideramos esta variável como possível variável explicativa da produção científica.

Segundo Long(1992), o gênero do cientista é um importante determinante da variação da produtividade científica. Diferenças existem em características pessoais, tais como habilidades, motivação, dedicação, nível educacional. Obrigações familiares e crianças afetam diferentemente as carreiras de homens e mulheres. Segundo este autor, as diferenças de gênero em termos de produtividade científica começam durante a graduação e continuam ao longo dos primeiros 17 anos da carreira do pesquisador. Rodgers e Maranto(1989) indicaram que os homens, em média, produzem significativamente mais publicações do

que as mulheres e são mais frequentemente citados. Entretanto Cole(1987) encontrou evidências de que a diferença na qualidade das publicações é menor do que a diferença na quantidade das mesmas. Estes mesmos pesquisadores, preocupados em explicar a diferença de produtividade entre os gêneros, encontraram resultados que mostraram que nem o estado civil nem o número de crianças conseguem explicar essa diferença. A posição hierárquica mais baixa e o menor salário, associado a menos recursos para a realização de pesquisas, podem ser fatores que explicam, em parte, essa produtividade diferencial entre homens e mulheres (Cole, 1987).

Estudando uma amostra de doutores em Bioquímica, Long(1992) relatou que enquanto as diferenças sexuais no número de publicações são crescentes na primeira década da carreira, diminuem na segunda. Uma significante proporção de mulheres não só mantêm sua produtividade, como a aumentam, ao passo que a média de produtividade masculina atinge níveis mais baixos.

Em nosso estudo queremos verificar se existe alguma relação entre o gênero e a produção científica. Consideramos que a variável **gênero** assume 1 quando masculino e 0 quando feminino.

Considerando que os professores encontram-se agrupados em centros acadêmicos, distintos principalmente pelas suas respectivas áreas do conhecimento foi então definido como 2° nível hierárquico os centros acadêmicos e, como variáveis candidatas a fazer parte do modelo neste nível: **Bolsista**, **área**, **CH** e **AH**, que serão explicadas a seguir.

A variável **Bolsista** é uma variável do 2°nível e indica a porcentagem de professores que são bolsistas do CNPq de cada centro. O objetivo da criação desta variável é saber se a produção científica de um docente pode ser influenciada pelo fato do mesmo pertencer a um centro em que a porcentagem de bolsista de produtividade é alta ou baixa.

Também foi criada uma variável categórica chamada **área** para separar os centros por área do conhecimento. A variável **área** é 1 para os centros CCEN,CTG e CIN; 2 para CCB e CCS; 3 para CFCH, CCSA; e 4 para CAC e CE. A finalidade da inclusão da

variável **área** é averiguar se a produção científica dos professores da UFPE é diferente por área do conhecimento na qual o professor está inserido.

Os docentes das instituições de ensino superior devem exercer por obrigação de sua função, atividades de pesquisa, ensino e extensão. É fato inquestionável que se o docente dedica muito tempo as suas atividades de ensino, a sua produção científica fica bastante prejudicada. Portanto, para medir o esforço docente nas atividades de ensino escolhemos duas variáveis, agregadas a nível de centro acadêmico, que são: **CH** que é a soma da carga horária total semanal nos cursos de graduação e pós-graduação de todos os docentes e **AH** que é a soma da quantidade de alunos matriculados multiplicado pela carga horária semanal nos cursos de graduação e pós-graduação de todos os docentes.

5.3.2 O Modelo

Na modelagem dos dados foram testados diversos modelos, incluindo uma variável por vez ou combinando as diversas variáveis disponíveis. Inicialmente foi considerado um modelo apenas com o intercepto para verificar se o mesmo poderia ser considerado como aleatório verificando o *valor-p*. Em seguida, foram incluídas variáveis explicativas apenas no nível 1 para verificar a significância das mesmas, sendo retiradas (uma por vez) aquelas que não apresentaram significância. Por fim, as variáveis significantes no nível 1 foram mantidas no modelo e aos parâmetros que poderiam ser considerados como aleatórios foram acrescentadas (uma após outra) variáveis explicativas no nível 2, mantendo-se aquela com nível de significância abaixo de 10%. Considerando os diversos modelos ajustados para explicar o índice de produção científica foi escolhido aquele estatisticamente significativo, cujas variáveis selecionadas estão na Tabela 5.3. Ainda na Tabela 5.3 têm-se a estatística descritiva das variáveis do modelo, na qual podemos observar que a média do tempo de serviço dos professores da UFPE é de 14,70 anos com um tempo de serviço máximo de 42 anos, quanto a porcentagem de bolsistas de produtividades por centro percebe-se que a porcentagem média é de 19% e que o centro com maior porcentagem possui 40% dos seus professores com bolsas de produtividade.

Variáveis	Média	Desvio Padrão	Mínimo	Máximo
Variável dependente				
Índice	211,14	213,67	0,00	2087
Variáveis independentes				
nível 1-professor (n=977)				
cod_idade (≤ 50 *anos* $= 1$)	0,60	0,09	0,00	1,00
T_Serv	14,70	10,21	1,00	42,00
gênero (masculino=1)	0,56	0,49	0,00	1,00
nível 2-centro (n=9)				
Bolsista (%)	0,19	0,09	0,04	0,40

Tabela 5.3: Estatísticas descritivas das variáveis do modelo

Além das variáveis da Tabela 5.3 também foi incluído no modelo a variável **área**. As estimativas apresentadas neste trabalho foram obtidas por meio do R *Project for Statistical Computing* (versão 2.6.2). O comando utilizado para estimar o modelo foi o *glmer* que necessita da biblioteca *lme4* para ser empregado.

O modelo selecionado para explicar o índice de produção científica do professor i do centro j foi:

Nível 1:

$$\eta_{ij} = \beta_{0j} + \beta_{1j}(cod_idade_{ij}) + \beta_{2j}(T_Serv_{ij}) + \beta_{3j}(gênero_{ij})$$

Nível 2:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}(Bolsista_j) + \gamma_{02}(área_j) + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}(área_j),
\end{aligned}$$

onde η_{ij} é o logaritmo do valor esperado do índice de produção científica do professor i do centro j, β_{qj} representa os coeficientes das variáveis do nível do professor do modelo e γ_{qp} representa o coeficiente das variáveis no nível do centro.

A Tabela 5.4 apresenta as estimativas dos parâmetros do modelo escolhido para explicar o índice de produção científica.

Analisando a Tabela 5.4 percebe-se que o tempo de serviço (**T_Serv**) está relacionado positivamente com o índice de produção científica. Para ser mais específico, a cada ano de serviço tem-se um acréscimo de $1,6\% = 100[exp(0,016) - 1]$ na média do índice de produção científica, mantendo todas as outras variáveis constantes. A variável **cod_idade** é estatisticamente significante e também se relaciona positivamente com o índice de produção científica. Docentes com idade menor ou igual a 50 anos apresentam em média o índice de produção científica maior em $36,8\%(1,368 = exp(0,314))$, fixando as outras variáveis. A variável **gênero** está relacionada negativamente com o índice de produção científica. Portanto, o professor sendo do gênero masculino ocorre um decréscimo na média do índice de produção científica de $7,8\% = 100[1 - exp(-0,081)]$, mantendo todas as outras variáveis constantes. Este fato se deve principalmente a área de artes e educação (CAC e CE) onde a participação do sexo feminino é maior e mais produtiva.

No nível 2(centro) a variável **Bolsista** se relaciona positivamente e fortemente com o logaritmo do índice de produção científica. Um aumento de 1% na porcentagem de bolsistas de produtividade em um centro afeta a média do índice de produção científica por um fator de $25,153 = exp(3,225)$, ou seja, 2415%. No modelo existe uma iteração entre as variáveis **gênero** e **área** e na Tabela 5.4 destaca-se que se o docente é do sexo masculino e pertence a área 2(CCB, CCS) há um aumento no valor médio do índice de produção científica de $10,4\% = 100[esp(0,099) - 1]$, se forem mantidas constantes as outras variáveis. Análise semelhante pode ser feita nos demais casos.

Efeitos Fixos	Estimativas	Erro Padrão	p-valor	exp(Estimativas)
Intercepto (γ_{00})	4,187	0,106	0,000	65,825
Bolsista (γ_{01})	3,225	0,144	0,000	25,153
área2 (γ_{02})	0,340	0,155	0,028	1,404
área3 (γ_{02})	-0,044	0,153	0,771	0,956
área4 (γ_{02})	0,275	0,156	0,077	1,316
cod_idade (γ_{10})	0,314	0,006	0,000	1,368
T_Serv (γ_{20})	0,016	0,000	0,000	1,016
gênero1 (γ_{30})	-0,081	0,008	0,000	0,922
gênero1:área2 (γ_{31})	0,099	0,011	0,000	1,104
gênero1:área3 (γ_{31})	0,257	0,013	0,000	1,293
gênero1:área4 (γ_{31})	-0,234	0,015	0,000	0,791
Efeito Aleatório	Desvio Padrão			
(u_{0j})	0.167			

Tabela 5.4: Estimativas dos parâmetros do modelo

Em resumo podemos então concluir que a significativa diferença na produção científica ocorre nos centros acadêmicos cuja quantidade relativa de bolsistas de produtividade do CNPq é grande, além disso com maior participação de docentes com idade até 50 anos. Em alguns centros as mulheres apresentam maior produção do que os homens, resultando diferenças estatísticas por gênero e área do conhecimento.

CAPÍTULO 6

Conclusões e Sugestões para Trabalhos Futuros

O objetivo geral deste trabalho foi o de criar um índice de produção científica dos docentes da UFPE independente da área e em seguida modelar este índice. Considerando que os professores encontram-se agrupados nos centros acadêmicos, distintos principalmente pelas suas respectivas áreas do conhecimento foi utilizado para modelar o índice de produção científica o MLGH com dois níveis em que o 1° nível é o nível do professor e o 2° nível é o do centro. Na modelagem a distribuição utilizada foi a Poisson, uma vez que os valores do índice são dados de contagem com um comportamento semelhante do da distribuição Poisson.

Na análise do índice de produção científica percebemos que o centro acadêmico com a maior média por docente é o CIN com o CCEN com a segunda maior média.

O modelo ajustado através do MLGH revelou que as variáveis significativas do nível 1 foram: o tempo de serviço que influencia positivamente no aumento da produção científica; a idade que de acordo com os resultados de Cole(1987) identifica que a produtividade cresce lentamente nas primeiras décadas de vida, alcançando seu pico na idade intermediária e decresce lentamente nas últimas décadas; a variável **gênero** que nos mostra que o índice de produção científica das mulheres é em média maior do que o dos homens.

Isso se dá devido ao fato de que em alguns centros acadêmicos(CAC, CCB, CE e CTG) a média dos índices de produção científica das mulheres é maior do que a dos homens.

Entre as variáveis do nível 2 a variável **Bolsista** mostrou uma forte relação com a produtividade, ou seja, como era de se esperar quanto maior a porcentagem de bolsistas de produtividade do CNPq no centro maior é o índice de produção docente. A iteração entre as variáveis **gênero** e **área** é interessante, pois quando consideramos o caso em que o docente é do sexo masculino e pertence a área 4(CAC, CE) temos um decréscimo no índice de produção, pois nesta área a média do índice de produção científica das mulheres é maior que o dos homens, como já referido.

Como sugestões para trabalhos futuros indicamos:

- acrescentar mais indicadores bibliométricos na construção do índice e adicionar variáveis no modelo de forma a melhor explicar o índice. Uma variável interessante a ser adicionada no modelo seria o tempo em que o docente terminou o doutorado;

- na modelagem através de análise de diagnóstico testar as suposições do modelo, bem como identificar pontos influentes;

- verificar a adequação de outras distribuições para esse conjunto de dados.

Referências Bibliográficas

[1] Araújo, Carlos, A. A. Bibliometria: evolução histórica. *Em Questão, Porto Alegre*, v. 12, n. 1, p. 11-32, jan./jun. 2006.

[2] Breslow, N.; Clayton, D. Approximate inference in generalized linear models. *Journal American Statistical Association*, v.88, p.9-25, 1993.

[3] Brookes, B. C. Biblio, sciento, informetrics? What are we talking about? In: EGGHE, L.; ROUSSEAU, R. (Ed.). *Informetrics* 89/90. Amsterdam : Elsevier, p. 31-43, 1990.

[4] Bryk, A.S.; Raudenbush, S.W. *Hierarchical Linear Models: Applications and data analysis methods*. 2 edition, Newbury Park, CA: Sage, 2002.

[5] Cole, J.R. *Fair science*: women in the scientific community. New York: Columbia University Press, 1987.

[6] Cordeiro, G.M.; McCullagh, P. Bias correction in generalized linear models- *J. R. Statistit. Soc. B*, v. 53, n.3, p. 629-643, 1991.

[7] Demétrio, C. G. B. *Modelos lineares generalizados na experimentação agronômica*. Porto Alegre: Universidade Federal do Rio Grande do Sul, 125p, 1993.

[8] Fonseca, E. N. Bibliografia estatística e bibliometria: uma reivindicação de prioridades. *Ciência da Informação*, Brasília, v. 2, n.1, p. 5-7, 1973.

[9] Fox, M. F. Publication productivity among scientist: a critical review. *Social of Studies*, v. 13, n.2, p.285-305, 1983.

[10] Garfield, E. Citation indexes to science: a new dimension in documentation through the association of ideas. *Science*, n.122, p.108-111, 1955.

[11] Goldstein, H . *Multilevel Statistical Models*, 1ªed. Internet London: Institute of Education, Multilevel Models Project, april 1999.

[12] Guedes. V.; Borschiver. S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica . *In Proceedings CINFORM - Encontro Nacional de Ciência da Informação VI*, Salvador - Bahia, 2005.

[13] Heyde, C. C. *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation*. New York: Springer-Verlag, 1997.

[14] J.E. Hirsch. An index to quantify an individual´s scintific research output, 2005.

[15] Lawani, S. M. Bibliometrics: its theoretical foundations, methods and applications. *Libri*, v. 31, n. 4, p. 294-315, 1981.

[16] Lehman, H.C. *Age and achievment*. Princeton: Princeton University Press, 1953.

[17] Lindley, D. V.; Smith, A. F. M. Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society*, Series B, n.34, p.1-41, 1972.

[18] Long, J. S. Measures of sex differences in scientific productivity. *Social Forces*, Chapel Hill, v.71, n.1, p.159-178, 1992.

[19] Macias-Chapula, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional.*Ciência da Informação, Brasília*, v. 27, n. 2, p. 134-140, maio/ago, 1998.

[20] Merton, R. K. The Matthew effect in science. *Science*, Washington, v.159, n.3810, p.56-63, 1968.

[21] McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2ªed. London: Chapman and Hall, 1989.

[22] Mugnaini, R.; Januzzi, P.; Quoiam, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. *Ciência da Informação*, Brasília, v. 33, n.2, 2004.

[23] Natis, L. *Modelos Lineares Hierárquicos*. Dissertação de mestrado defendida no Instituto de Matemática e Estatística da Universidade de São Paulo, 2000.

[24] Nelder, J.A.; Wedderburn, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society*, A 135, 370-384, 1972.

[25] Nicholas, D.; Ritchie, M. *Literature and Bibliometrics*. London: Clive Bingley, 1978.

[26] Okubo, Y. *Bibliometric Indicators and Analysis of Research Systems: methods and examples*. Paris: OCDE/GD, 1997.

[27] Paula, A.G. *Modelos de Regressão com Apoio Computacional*, IME-USP, jun, 2004.

[28] Pinheiro, S.M.C. *Modelo Linear Hierárquico: Um método alternativo para análise de desempenho escolar*. Dissertação de mestrado defendida no Departamento de Estatística da Universidade Federal de Pernambuco, 2005.

[29] Rodgers, R.C.; Maranto, C.L. Causal models of publishing productivity in psychology, *Journal of Applied Psychology*, v. 74, n.4, p.636-649, 1989.

[30] Schall, R. Estimation in generalized linear models with random effects. *Biometrika*, v.40, p.719-727, 1991.

[31] Sengupta, I. N. Bibliometrics, informetrics, scientometrics and librametrics: an overview. *Libri*, v. 42, n.2, p. 99-135, 1992.

[32] Spinak, E. *Diccionario Enciclopédico de Bibliometria, cienciometria e informetria*. Montevideo, 1996. 245 p.

[33] Spinak, E. Indicadores cienciométricos. *Ciência da Informação*, Brasília, v. 27, n. 2, p. 141-148, maio/ago, 1998.

[34] Sullivan, L.M., Duke, K. & Losina, E. Tutorial in biostatistics an introduction to hierarchical linear modelling. *Statistics in Medicine*, v.18, p,855-888, 1999.

[35] Tangue-Sutcliffe, Jean. An introduction to informetrics. *Information processing and management, Oxford*, v. 28, n. 1, p. 1-3, 1992.

[36] Van Raan, A. F. J. Scientometrics: state-of-art. *Scientometrics*, v. 38, n. 1, p. 205-218, 1997.

[37] Vanti, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ciência da Informação, Brasília*, v. 31, n. 2, p. 152-162, maio/ago, 2002.

[38] Zorn, C. J. W. *GEE Models of Judicial Behavior*. Atlanta, 1998. Trabalho não publicado.

Apêndice

Programas Utilizados

*****Modelo*****

```
bd<-read.table("F:/dissertação/dados2/ava4.dat",header=T)
 mod1<- glmer(Indice ~ Bolsista  + area + genero + cod_idade +
 T_serv + genero*area +(1 | n_centro), family=poisson, data=bd)
```

*****Histograma*****

```
bd<-read.table("F:/dissertação/dados2/ava4.dat",header=T)
hist(bd\$Indice, breaks=40, main="Histograma - Índice de Produção", xlab="Índice",
 ylab="Frequências Absolutas", col="blue")
```