



Universidade Estadual de Campinas

Instituto de Matemática, Estatística e  
Computação Científica - IMECC  
Departamento de Estatística



# MODELOS MULTINOMIAIS MULTIVARIADOS APLICADOS EM SEQÜÊNCIAS DE DNA

Dissertação de Mestrado

Beatriz Castro Dias Cuyabano

Orientadora: **Profa. Dra. Hildete Prisco Pinheiro**

CAMPINAS

Fevereiro/2011

# Modelos Multinomiais Multivariados Aplicados em Sequências de DNA

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Beatriz Castro Dias Cuyabano, aprovada pela Comissão Julgadora.

Campinas 25 de fevereiro de 2011



Prof. Dra. Hildete Prisco Pinheiro  
Orientadora

Banca Examinadora:

- Profa. Dra. Hildete Prisco Pinheiro (IMECC - Unicamp) - Orientadora
- Prof. Dr. Víctor Hugo Lachos Dávila (IMECC - Unicamp)
- Prof. Dr. Juvêncio Santos Nobre (DEMA - UFC)

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da UNICAMP, como requisito parcial para obtenção do Título de Mestre em Estatística.

**FICHA CATALOGRÁFICA ELABORADA PELA  
BIBLIOTECA DO IMECC DA UNICAMP**  
Bibliotecária: Maria Fabiana Bezerra Müller – CRB8 / 6162

Cuyabano, Beatriz Castro Dias  
C99m Modelos multinomiais multivariados aplicados em seqüências de  
DNA/Beatriz Castro Dias Cuyabano-- Campinas, [S.P. : s.n.], 2011.

Orientador : Hildete Prisco Pinheiro  
Dissertação (mestrado) - Universidade Estadual de Campinas,  
Instituto de Matemática, Estatística e Computação Científica.

1.Bahadur, Representação de. 2.DNA. 3.Modelos de regressão.  
I. Pinheiro, Hildete Prisco. II. Universidade Estadual de Campinas.  
Instituto de Matemática, Estatística e Computação Científica. III. Título.

Título em inglês: Multivariate multinomial models applied to DNA sequences

Palavras-chave em inglês (Keywords): 1. Bahadur representation. 2.DNA. 3.Reggression models.

Área de concentração: Estatística

Titulação: Mestre em Estatística

Banca examinadora: Profª. Dra. Hildete Prisco Pinheiro (IMECC – UNICAMP)  
Prof. Dr. Víctor Hugo Lachos Dávila (IMECC - UNICAMP)  
Prof. Dr. Juvêncio Santos Nobre (UFC)

Data da defesa: 25/02/2011

Programa de Pós-Graduação: Mestrado em Estatística

**Dissertação de Mestrado defendida em 25 de fevereiro de 2011 e aprovada**

**Pela Banca Examinadora composta pelos Profs. Drs.**



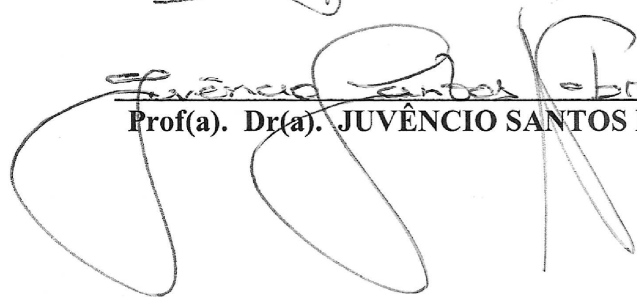
---

**Prof(a). Dr(a). HILDETE PRISCO PINHEIRO**



---

**Prof(a). Dr(a). VÍCTOR HUGO LACHOS DÁVILA**



---

**Prof(a). Dr(a). JUVÊNCIO SANTOS NOBRE**

*Para Amanda Lucas Gimeno (in memoriam), cuja amizade foi de grande influência e  
motivação para a realização do mestrado.*

# Agradecimentos

Agradeço a meus pais Tuca e Sérgio, e meu irmão Thiago que sempre me apoiaram e incentivaram para que o mestrado, e muitos outros objetivos, se tornassem possíveis. A minha mãe em particular por todas as revisões de texto durante este projeto.

A minha orientadora, Professora Hildete Pinheiro, por toda a atenção, amizade e dedicação, não somente no desenvolvimento do trabalho do mestrado, como também no desenvolvimento de projetos futuros, e ao Professor Alúcio Pinheiro, pela grande contribuição nas soluções dos problemas computacionais, e novas propostas para análises dos dados.

Aos amigos da pós-graduação, e do IMECC, que estiveram juntos nas aulas, nas horas de estudos, nos cafés ou sucos e nos momentos de lazer e distração, fazendo do mestrado e dos dias na UNICAMP sempre melhores.

Aos Professores Victor Hugo Lachos e Juvêncio dos Santos Nobre pela participação na banca de defesa do mestrado.

À Lori Cristina Grandin, cuja dissertação de mestrado serviu de base inicial para este trabalho, e que gentilmente forneceu o banco de dados utilizado.

À CAPES, pelo suporte financeiro fundamental, sem o qual a realização deste projeto não seria possível.

# Resumo

Modelos Multivariados são propostos para descrever a frequência de códons em seqüências de DNA, bem como a ordem e frequência em que as bases nitrogenadas se apresentam em cada códon, considerando a dependência entre as bases dentro do códon. Modelos logísticos regressivos são utilizados com diferentes estruturas de dependência entre as posições do códon. Também, modelos baseados em uma extensão da representação de Bahadur para o caso multinomial são propostos para explicar dados multinomiais correlacionados. Uma aplicação desses modelos para o gene NADH4 do genoma mitocondrial humano é apresentada, e comparações desses modelos são feitas a partir de diferentes critérios como AIC, BIC e validação cruzada. Por fim, uma breve análise de diagnósticos é realizada para os modelos logísticos regressivos.

# Abstract

Multivariate models are proposed to describe the codons frequencies in DNA sequences, as well as the order and frequency that nucleotide bases have in each codon, considering the dependence among the bases inside a codon. Logistic regressive models are used with different structures of dependence among the three positions in a codon. Also, models based on a multinomial extension of the Bahadur's representation are proposed to explain correlated multinomial data. An application of these models to the NADH4 gene from human mitochondrial genome is presented, and model comparisons among them are done by different criteria such as AIC, BIC and cross validation. At last, a brief diagnostic analysis is done upon the logistic regressive models.



# Sumário

<b>Lista de Figuras</b>	<b>xvii</b>
<b>Lista de Tabelas</b>	<b>xxi</b>
<b>1 Introdução</b>	<b>1</b>
1.1 Biologia Molecular . . . . .	3
1.2 Banco de Dados . . . . .	7
1.3 Análise Exploratória . . . . .	9
<b>2 Modelos Binomiais Multivariados</b>	<b>13</b>
2.1 Modelos Lineares Generalizados . . . . .	14
2.1.1 Distribuição de Bernoulli e Binomial . . . . .	16
2.2 Funções de Ligação mais Comuns em Modelos Binomiais . . . . .	16
2.2.1 Logito . . . . .	17
2.2.2 Probitto . . . . .	17
2.2.3 Log-Log Complementar . . . . .	17
2.3 Modelos Logísticos Regressivos . . . . .	18
2.3.1 Modelo Independente . . . . .	19
2.3.2 Modelo Igualmente Preditivo . . . . .	20
2.3.3 Estrutura Markoviana de Primeira Ordem . . . . .	21
2.3.4 Modelo Aditivo . . . . .	21
2.3.5 Gradiente e Informação de Fisher . . . . .	22

---

2.4	Modelo Baseado na Representação de Bahadur . . . . .	25
2.4.1	Gradiente e Informação de Fisher . . . . .	29
<b>3</b>	<b>Modelos Multinomiais Multivariados</b>	<b>31</b>
3.1	Modelos Logísticos Regressivos . . . . .	32
3.1.1	Modelo Independente . . . . .	34
3.1.2	Modelo Igualmente Preditivo . . . . .	35
3.1.3	Estrutura Markoviana de Primeira Ordem . . . . .	36
3.1.4	Modelo Aditivo . . . . .	36
3.1.5	Gradiente e Informação de Fisher . . . . .	37
3.2	Modelo Baseado na Representação de Bahadur . . . . .	41
3.2.1	Modelo de Dependência de Locação . . . . .	43
3.2.2	Modelo de Dependência de Transição . . . . .	44
3.2.3	Modelo de Dependência de Semi-Locação e Transição . . . . .	44
3.2.4	Modelo de Dependência de Locação e Transição . . . . .	45
3.2.5	Método de Estimação dos Parâmetros . . . . .	45
3.2.6	Gradiente e Informação de Fisher . . . . .	46
<b>4</b>	<b>Medidas de Ajuste dos Modelos</b>	<b>49</b>
4.1	Soma de Quadrado dos Erros (SQE) . . . . .	49
4.2	Critério de Informação de Akaike (AIC) . . . . .	49
4.3	Critério de Informação Bayesiano (BIC) . . . . .	50
4.4	Função Desvio . . . . .	50
4.5	Validação Cruzada . . . . .	50
4.5.1	K-dobras . . . . .	51
4.5.2	<i>Hold-out</i> . . . . .	51
4.5.3	<i>Leave-one-out</i> . . . . .	51
4.5.4	K-dobras repetido . . . . .	52
4.6	Teste da Razão de Verossimilhança . . . . .	52
4.7	Teste de Wald . . . . .	52

---

4.8	Teste de Escore . . . . .	53
<b>5</b>	<b>Aplicação e Resultados</b>	<b>55</b>
5.1	Implementação Computacional . . . . .	55
5.2	Modelos Binomiais . . . . .	56
5.3	Modelos Multinomiais . . . . .	57
<b>6</b>	<b>Análise de Diagnóstico dos Modelos</b>	<b>67</b>
6.1	Modelos Logísticos Regressivos . . . . .	68
6.1.1	Modelo Independente . . . . .	71
6.1.2	Modelo Iguamente Preditivo . . . . .	71
6.1.3	Estrutura Markoviana de Primeira Ordem . . . . .	72
6.1.4	Modelo Aditivo . . . . .	73
6.2	Resultados . . . . .	73
<b>7</b>	<b>Considerações Finais</b>	<b>83</b>
	<b>Referências Bibliográficas</b>	<b>87</b>
	<b>Apêndice 1</b>	<b>91</b>
	<b>Apêndice 2</b>	<b>97</b>
	<b>Apêndice 3</b>	<b>103</b>
	<b>Apêndice 4</b>	<b>107</b>

# Lista de Figuras

1.1	Estrutura do DNA (Imagem: Wikipedia, Autor: Michael Ströck) . . . . .	6
1.2	Estrutura Química do DNA (Imagem: Wikipedia) . . . . .	6
1.3	Proporção das Bases Nitrogenadas em Cada Posição dos Códon . . . . .	9
5.1	Valores observados <i>versus</i> estimados para os códon do modelo logístico regressivo aditivo com ligação logito . . . . .	58
5.2	Valores observados <i>versus</i> estimados para os códon do modelo baseado na representação de Bahadur com ligação logito . . . . .	58
5.3	Valores observados <i>versus</i> estimados do modelo logístico regressivo aditivo	60
5.4	Valores observados <i>versus</i> estimados do modelo baseado na representação de Bahadur com dependência de semi-locção e transição . . . . .	60
5.5	Validação Cruzada dos Modelos Logísticos Regressivos . . . . .	64
5.6	Validação Cruzada dos Modelos Baseados na Representação de Bahadur	65
6.1	Diagnóstico do Modelo Independente . . . . .	74
6.2	Diagnóstico do Modelo Igualmente Preditivo . . . . .	75
6.3	Diagnóstico da Estrutura Markoviana de Primeira Ordem . . . . .	76
6.4	Diagnóstico do Modelo Aditivo . . . . .	77

# Lista de Tabelas

1.1	Bases Nitrogenadas . . . . .	4
1.2	Código Genético Mitocondrial para Mamíferos - Aminoácidos . . . . .	5
1.3	Total das Bases nas Posições 1 e 2 do Códon . . . . .	9
1.4	Total das Bases nas Posições 1 e 3 do Códon . . . . .	10
1.5	Total das Bases nas Posições 2 e 3 do Códon . . . . .	10
1.6	Testes Chi-Quadrado de Independência entre as Posições . . . . .	11
3.1	Codificação das Bases Nitrogenadas . . . . .	32
5.1	Medidas dos Modelos Binomiais . . . . .	57
5.2	Medidas dos Modelos Multinomiais Regressivos . . . . .	59
5.3	Medidas dos Modelos Multinomiais Baseados na Representação de Bahadur . . . . .	59
5.4	Estimativas dos Parâmetros do Modelo Logístico Regressivo Aditivo . . . . .	61
5.5	Estimativas dos Parâmetros do Modelo Baseado na Representação de Bahadur de Dependência de Semi-Localização e Transição . . . . .	62
5.6	Resultados da Validação Cruzada dos Modelos Logísticos Regressivos . . . . .	63
5.7	Resultados da Validação Cruzada dos Modelos Baseados na Representação de Bahadur . . . . .	63
5.8	Testes dos Parâmetros do Modelo Aditivo e do Modelo de Semi-Localização & Transição . . . . .	64
6.1	Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Independente . . . . .	78

---

6.2	Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Iguamente Preditivo . . . . .	79
6.3	Estimativas dos Parâmetros Retirando Observações Discrepantes da Estrutura Markoviana . . . . .	80
6.4	Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Aditivo . . . . .	81
7.1	Número de Parâmetros dos Modelos Multinomiais Multivariados para $K$ Posições Dependentes . . . . .	84
7.2	Total de Parâmetros de Dependência dos Modelos Multinomiais Multivariados Conforme o Número $K$ de Posições Dependentes Aumenta . . . . .	85
7.3	Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Logito . . . . .	98
7.4	Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Probito . . . . .	99
7.5	Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Log-Log Complementar . . . . .	100
7.6	Probabilidades Estimadas dos Modelos Multinomiais Multivariados Logísticos Regressivos . . . . .	101
7.7	Probabilidades Estimadas dos Modelos Multinomiais Multivariados Baseados na Representação de Bahadur . . . . .	102
7.8	Parâmetros Estimados dos Modelos Logístico Regressivos Independentes Binomiais Multivariados . . . . .	103
7.9	Parâmetros Estimados dos Modelos Logístico Regressivos Iguamente Preditivos Binomiais Multivariados . . . . .	103
7.10	Parâmetros Estimados dos Modelos Logístico Regressivos com Estrutura Markoviana de Primeira Ordem Binomiais Multivariados . . . . .	104
7.11	Parâmetros Estimados dos Modelos Logístico Regressivos Aditivos Binomiais Multivariados . . . . .	104

---

7.12	Parâmetros Estimados dos Modelos Baseados na Representação de Bahadur Binomiais Multivariados . . . . .	104
7.13	Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Independente . . . . .	105
7.14	Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Igualmente Preditivo . . . . .	105
7.15	Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Estrutura Markoviana de Primeira Ordem . . . . .	105
7.16	Parâmetros Estimados do Modelo Baseado na Representação de Bahadur de Dependência de Locação . . . . .	105
7.17	Parâmetros Estimados do Modelo Baseado na Representação de Bahadur de Dependência de Transição . . . . .	106
7.18	Parâmetros Estimados do Modelo Baseado na Representação de Bahadur de Dependência de Locação e Transição . . . . .	106

# Capítulo 1

## Introdução

O seqüenciamento genético tem cada vez mais atraído a atenção e sido tema freqüente de pesquisas científicas por todo o mundo. Esse fato é facilmente compreensível, uma vez que a quantidade de informações que podem ser obtidas a partir de um único gene<sup>1</sup> é capaz de responder inúmeras perguntas da ciência. Além disso, avanços computacionais, o fácil acesso aos dados e o crescente número de pesquisas envolvendo esse assunto favorecem o estudo e aperfeiçoamento de técnicas voltadas para essas análises.

Neste trabalho, o foco principal é a análise da freqüência dos códons<sup>2</sup> presentes no gene NADH4 (dehidrogenase) do genoma mitocondrial humano, e também a seqüência de nucleotídeos - Timina (T), Citosina (C), Adenina (A) e Guanina (G) - que formam esses códons, bem como suas classificações proteicas, em *pirimidina* (T,C) ou *purina* (A,G). A base para esse estudo é a *seqüência de referência de Cambridge* (SRC), primeiramente publicada em 1981 ([Anderson et al., 1981](#)) e revisada em 1999 ([Andrews et al., 1999](#)), aceita mundialmente como referência de seqüência de DNA mitocondrial humano (DNAmit).

Modelos de respostas binomiais e multinomiais são propostos, gerados, analisados e comparados, através de métodos estatísticos, para 30 seqüências do gene NADH4, dentre elas a SRC, obtidas no site do *National Center for Biotechnology Information*

---

<sup>1</sup>Segmento de DNA que contém informações para a síntese de uma ou mais proteínas.

<sup>2</sup>Tripla de nucleotídeos adjacentes.



---

(NCBI, 2010), a fim de se obter um modelo mais parcimonioso para tal seqüenciamento.

A importância de estudos estatísticos sobre esse tema se deve principalmente ao interesse de pesquisadores em compreender e prever estruturas genéticas. A complexidade do DNA exige modelos com altos níveis de detalhamento, como os modelos logísticos regressivos (Bonney et al., 1994), que analisam a dependência entre as três posições em um códon fazendo uso das posições anteriores como covariáveis para os modelos logísticos das seguintes.

Após vasto estudo sobre os modelos logísticos regressivos, pesquisas levaram à representação de Bahadur (Bahadur, 1961), que define uma função de probabilidade conjunta para dados binários correlacionados. Há também uma discussão sobre a representação de Bahadur como alternativa aos modelos logísticos (Cox, 1972) para se escrever a probabilidade conjunta de dados binários multivariados diretamente em termos das probabilidades ao invés de log-probabilidades, e Zhao e Prentice (1990) fazem uso da representação de Bahadur como uma caso particular do modelo quadrático exponencial para dados binários correlacionados.

Tal representação também é encontrada na literatura em estudos de respostas dicotômicas permutáveis para indivíduos classificados em grupos (Stefanescu e Turnbull, 2003), e para estudos longitudinais, em que uma mesma variável resposta é obtida para tempos diferentes em uma mesma unidade (Fitzmaurice et al., 1993; Fitzmaurice, 1995).

Além disso há o uso da representação de Bahadur na modelagem de dados longitudinais (Parzen et al., 2009) com abordagem de autocorrelação, devido à característica de série temporal dos dados e utilizando métodos de estimação que envolvem modelos autorregressivos.

A representação de Bahadur também já foi utilizada para analisar mutações em estruturas genéticas (Pinheiro et al., 1999), portanto o uso dessa representação, e de sua expansão para dados multinomiais correlacionados, é pertinente para a modelagem da estrutura de dependência em um códon.

Em sua primeira publicação (Bonney, 1986), os modelos logísticos regressivos foram

aplicados a dados de família, e mais tarde aplicados em apenas uma seqüência de DNA (Bonney et al., 1994). Grandin (2006) aborda a aplicação em diversas seqüências. A representação de Bahadur para o caso de respostas multinomiais, apesar de sugerida e brevemente descrita por Bahadur, não consta em trabalhos como utilizada para modelar dados genéticos.

Ainda na introdução há uma descrição do banco de dados utilizado, bem como alguns detalhes genéticos relacionados a esse banco de dados e possíveis covariáveis, também sugeridas por Bonney para os modelos. Os Capítulos 2 e 3 apresentam respectivamente os modelos teóricos para dados binomiais e multinomiais multivariados, bem como problemas e soluções da estimação dos parâmetros por máxima verossimilhança (EMV).

Algumas medidas de ajuste e técnicas de comparação de modelos são introduzidas no Capítulo 4, e o Capítulo 5 traz a aplicação dos modelos propostos no banco de dados apresentado. Em seguida, o Capítulo 6 aborda uma breve análise de diagnóstico dos modelos logísticos regressivos multinomiais. Por fim, o Capítulo 7 apresenta uma discussão, a partir dos resultados obtidos, sobre as vantagens e desvantagens dos modelos logísticos regressivos e dos modelos baseados na representação de Bahadur.

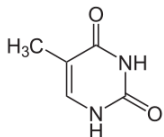
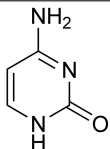
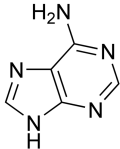
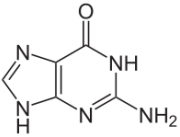
## 1.1 Biologia Molecular

O ácido desoxirribonucleico (DNA, do inglês *desoxyribonucleic acid*) é um ácido encontrado dentro do núcleo das células de todos os seres vivos, com a exceção de alguns vírus, com a função de armazenar as informações necessárias para o funcionamento e características destes seres vivos, como uma “instrução” de como as células e proteínas, por exemplo, devem ser sintetizadas. Sua primeira publicação, feita por Watson e Crick (1953), define o DNA como uma dupla hélice, diferente da idéia que havia até então sobre os ácidos nucleicos, de que estes eram compostos por uma estrutura de três tiras.

O DNA pode ser subdividido em cromossomos, e cada cromossomo em genes, que são as unidades de hereditariedade passadas nas gerações e contém as informações que

influenciam em características particulares de cada organismo. O código genético de cada gene é determinado por uma seqüência de nucleotídeos (ou bases nitrogenadas), sendo eles Timina (T), Citosina (C), Adenina (A) e Guanina (G), cujas breves descrições químicas se encontram na Tabela 1.1. O DNA humano é composto por um total de 23 pares de cromossomos, e o número de genes contidos nesses cromossomos ainda não é conhecido, mas estima-se que seja entre 20 e 25 mil.

Tabela 1.1: Bases Nitrogenadas

Nucleotídeo	Fórmula Química	Molécula
Timina	$C_5H_6N_2O_2$	
Citosina	$C_4H_5N_3O$	
Adenina	$C_5H_5N_5$	
Guanina	$C_5H_5N_5O$	

Cada grupo de três nucleotídeos adjacentes configuram um códon, e como há quatro nucleotídeos possíveis, existem 64 diferentes códons. Os cdons, por sua vez, determinam algum dentre os 20 aminoácidos existentes na cadeia proteica. Mais de um códon pode determinar o mesmo aminoácido e aqueles que determinam um mesmo aminoácido são chamados de códons sinônimos. Há também códons que não sintetizam aminoácidos, mas indicam o término da síntese proteica (fim de um gene), ou seja, são códons de parada e considerados não-efetivos. Além disso, o aminoácido Metionina indica o início

da síntese proteica (início de um gene). A Tabela 1.2 mostra a relação entre códons e aminoácidos.

Tabela 1.2: Código Genético Mitochondrial para Mamíferos - Aminoácidos

Aminoácidos	Sigla	Códons Sinônimos
Alanina	Ala/A	GCT, GCC, GCA, GCG
Arginina	Arg/R	CGT, CGC, CGA, CGG
Asparagina	Asn/N	AAT, AAC
Ácido Aspártico	Asp/D	GAT, GAC
Cisteína	Cys/C	TGT, TGC
Ácido Glutâmico	Glu/E	GAA, GAG
Glutamina	Gln/Q	CAA, CAG
Glicina	Gly/G	GGT, GGC, GGA, GGG
Histidina	His/H	TAT, CAC
Isoleucina	Ile/I	ATT, ATC
Leucina	Leu/L	TTA, TTG, CTT, CTC, CTA, CTG
Lisina	Lys/K	AAA, AAG
Metionina	Met/M	ATA, ATG
Fenilalanina	Phe/F	TTT, TTC
Prolina	Pro/P	CCT, CCC, CCA, CCG
Serina	Ser/S	TCT, TCC, TCA, TCG, AGT, AGC
Treonina	Thr/T	ACT, ACA, ACG, ACC
Triptofano	Trp/W	TAG, TGG
Tirosina	Tyr/Y	TAT, TAC
Valina	Val/V	GTT, GTC, GTA, GTG
Códons de Parada	<i>Ter</i>	TAA, TAG, AGA, AGG

O DNA não é composto por uma única molécula, mas sim por um par de moléculas, como duas fitas, fortemente ligadas, no formato de uma dupla hélice. Cada uma dessas

fitas é composta por uma seqüência de bases nitrogenadas, unidas por um esqueleto de fosfato e resíduos de açúcar (fosfato-desoxirribose), e a dupla hélice é unida por pontes de hidrogênio, respeitando a complementariedade das bases, em que uma *pirimidina* (T ou C) faz uma ligação única sempre a uma *purina* (A ou G), assim, T se une somente a A com duas pontes de hidrogênio, e C se une somente a G com três pontes de hidrogênio.

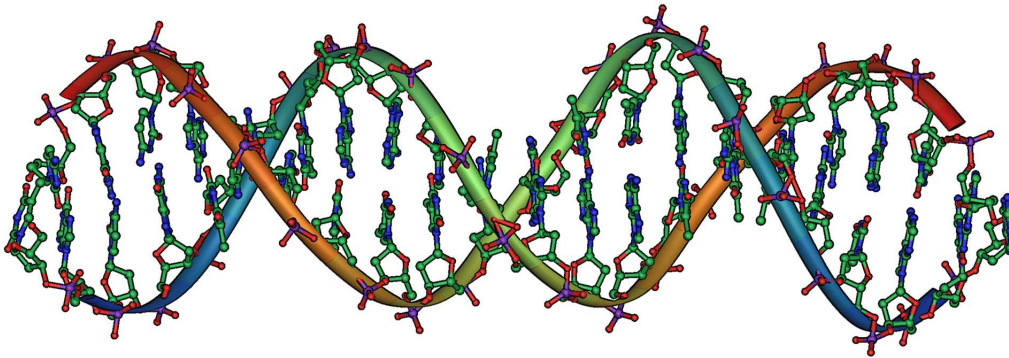


Figura 1.1: Estrutura do DNA (Imagem: Wikipedia, Autor: Michael Ströck)

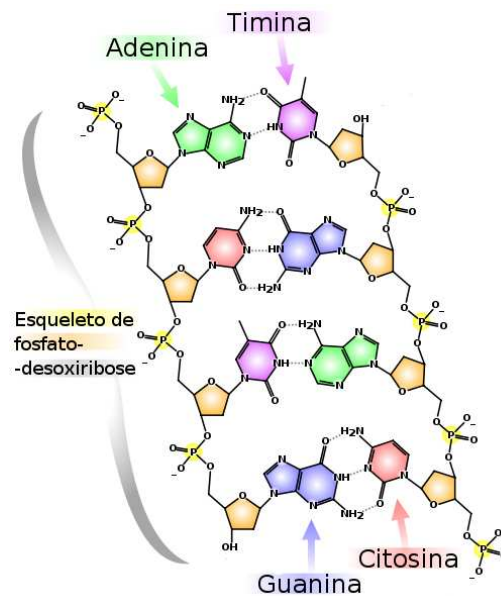


Figura 1.2: Estrutura Química do DNA (Imagem: Wikipedia)

A informação genética contida no DNA é copiada pelo ácido ribonucleico (RNA) mensageiro, em um processo denominado transcrição. Neste processo, o RNA cria

uma réplica com uma única fita complementar àquela copiada do DNA, e diferente da replicação do DNA, o RNA possui a base Uracila (U) ao invés da Timina, com a Adenina como complementar. Por fim, essa cópia de RNA é decodificada por um ribossomo em um processo denominado tradução, e o RNA mensageiro é emparelhado a um RNA de transferência, que é responsável pela transmissão do código genético.

Algumas leituras são recomendadas para mais detalhes e aprofundamento a respeito de genética e biologia molecular, como por exemplo [Alberts et al. \(2002\)](#), e para aplicações estatísticas em genética humana e biologia molecular, [Reilly \(2009\)](#).

## 1.2 Banco de Dados

O banco de dados utilizado para as análises consiste de 30 seqüências de DNA obtidas no site do NCBI, com as seguintes características:

- 1 é a SRC
- 7 têm a doença de Leber
- 4 têm o Mal de Alzheimer
- 1 tem Diabetes
- 8 têm o Mal de Parkinson
- 1 é obeso
- 8 são normais

Todas essas seqüências se apresentam codificadas pelos nucleotídeos, ou seja, são seqüências na forma (...)atgctaaaac(...). Além das bases nitrogenadas do gene NADH4, foram consideradas também três covariáveis, introduzidas por [Bonney et al. \(1994\)](#): AARISK como uma medida do risco de mutação em um aminoácido, AVDIST como uma medida para o quão típico é um aminoácido e TSCORE que mede o número de mudanças únicas em um único nucleotídeo que podem transformar o códon em um códon de parada (por exemplo, o códon TTA tem TSCORE 1, pois ao mudar o segundo

T para A obtém-se o códon de parada TAA, já o códon TGA tem TSCORE 2, pois ao mudar G para A, obtém-se o códon de parada TAA, ou ao mudar o primeiro T para A, obtém-se o códon de parada AGA). Em geral AARISK e TSCORE assumem valores diferentes para códonos sinônimos.

As covariáveis AARISK e AVDIST são obtidas a partir das seguintes propriedades químicas dos aminoácidos: composição ( $c$ ), polaridade ( $p$ ) e volume molecular ( $v$ ); também por  $\alpha$ ,  $\beta$  e  $\gamma$  que são os quadrados do inverso das médias, respectivamente da composição, polaridade e volume molecular de todos os 20 aminoácidos (Grantham, 1974). AARISK é a média ponderada das distâncias entre um aminoácido e os demais, e AVDIST é a média das distâncias entre um aminoácido e os demais, sem ponderação, sendo que, quanto menor AVDIST, mais típico é o aminoácido. As distâncias são dadas pela equação,

$$D_{ij} = \sqrt{[\alpha (c_i - c_j)^2 + \beta (p_i - p_j)^2 + \gamma (v_i - v_j)^2]}. \quad (1.1)$$

Outro fato importante é que as covariáveis não têm valores atribuídos aos códonos de parada, uma vez que TSCORE mede o número de mudanças únicas em um único nucleotídeo que podem transformar o códon em um códon de parada. Não há sentido em aplicar tal medida em um códon que originalmente configura parada. Portanto, os modelos que incluem covariáveis contam com apenas 60 frequências, e não 64, após excluídos os quatro códonos de parada.

Todas as 30 seqüências do gene NADH4 que compõem a amostra são consideradas independentes e cada uma possui um total de 460 códonos quando considerado o último códon que é de parada; há então 459 códonos efetivos, também considerados independentes entre si, totalizando uma amostra com  $N = 13770$  códonos efetivos a serem analisados.

## 1.3 Análise Exploratória

Uma breve análise exploratória foi realizada sobre o banco de dados, analisando a distribuição das bases nitrogenadas para cada posição dos códons, conforme mostra a Figura 1.3. É visível, nos gráficos de barra, a diferença na proporção dos nucleotídeos em cada posição.

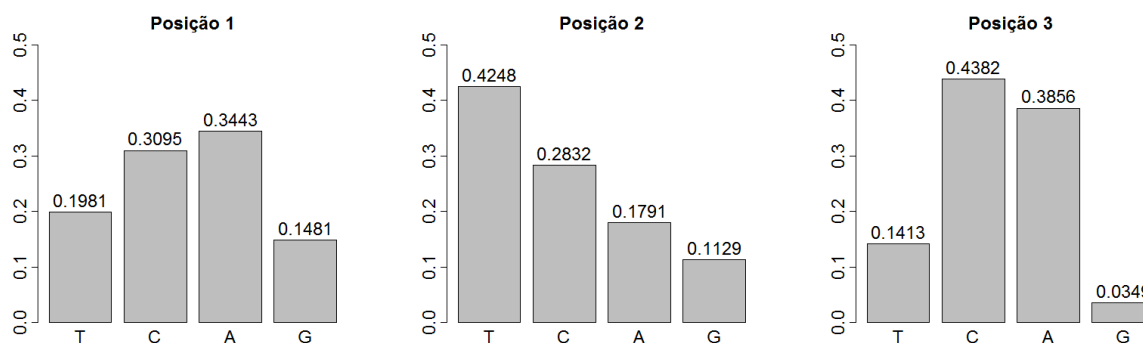


Figura 1.3: Proporção das Bases Nitrogenadas em Cada Posição dos Códons

Outra análise feita sobre os dados, foi a respeito da transição (mudança) de uma base para outra na posição seguinte, ou seja, quais as bases das posições 1 e 2 do códon, e para quais bases mudam nas posições 2 e 3. As Tabelas 1.3, 1.4 e 1.5 mostram o distribuição do total de bases nas três posições, comparando-as entre si.

Tabela 1.3: Total das Bases nas Posições 1 e 2 do Códon

		Posição 2				Total
		T	C	A	G	
Posição 1	T	868	990	390	480	2728
	C	2612	690	696	264	4262
	A	1980	1440	1021	300	4741
	G	390	780	359	510	2039
Total		5850	3900	2466	1554	13770



Tabela 1.4: Total das Bases nas Posições 1 e 3 do Códon

		Posição 3				Total
		T	C	A	G	
Posição 1	T	512	1228	898	90	2728
	C	407	1873	1834	148	4262
	A	816	2094	1682	149	4741
	G	211	839	895	94	2039
Total		1946	6034	5309	481	13770

Tabela 1.5: Total das Bases nas Posições 2 e 3 do Códon

		Posição 3				Total
		T	C	A	G	
Posição 2	T	1052	2068	2462	268	5850
	C	645	1815	1381	59	3900
	A	129	1437	839	61	2466
	G	120	714	627	93	1554
Total		1946	6034	5309	481	13770

Em seguida, testes chi-quadrado de Pearson foram aplicados aos dados para verificar independência das posições. Como todas as tabelas são  $4 \times 4$ , as estatísticas  $Q$  dos testes possuem distribuição assintótica  $\chi_9^2$  sob a hipótese nula. Portanto, com nível de significância  $\alpha = 0,05$ , se  $Q > 16,92$ , rejeita-se a hipótese de independência entre as posições. Os resultados dos testes encontram-se na Tabela 1.6, e percebe-se que as posições não são independentes com nível de significância de 5%.

Tabela 1.6: Testes Chi-Quadrado de Independência entre as Posições

<b>Posições</b>	<b>Q</b>	<b>g.l.</b>	<b>p-valor</b>	<b>Resultado</b>
1 e 2	1749,385	9	$< 2,2 \times 10^{-16}$	não independentes
1 e 3	242,257	9	$< 2,2 \times 10^{-16}$	não independentes
2 e 3	624,197	9	$< 2,2 \times 10^{-16}$	não independentes

## Capítulo 2

# Modelos Binomiais Multivariados

Seja o  $i$ -ésimo códon representado pelo vetor  $\mathbf{Y}_i = (Y_{1i}, Y_{2i}, Y_{3i})$ , em que  $Y_{ki}$  (tal que  $k = 1, 2, 3$  é a posição no códon) assume uma dentre as bases nitrogenadas T, C, A ou G. Considerando apenas os 60 códons efetivos, conforme explicado anteriormente na introdução na seção sobre o banco de dados, a distribuição deles nas seqüências de NADH4 é multinomial, ou seja,  $(\mathbf{Y}_1, \dots, \mathbf{Y}_{60}) \sim M(N, p_1, \dots, p_{60})$ , tal que  $p_i = P(\text{observar o códon } \mathbf{Y}_i)$  é a probabilidade de que um códon selecionado dentro de uma amostra seja igual ao códon representado pelo vetor  $\mathbf{Y}_i$ , para todo  $i = 1, \dots, 60$ .

Olhando para uma amostra composta por  $N$  códons, a probabilidade de que cada códon  $\mathbf{Y}_i$  seja observado  $n_i$  vezes dentro dessa amostra ( $\#\mathbf{Y}_i = n_i$ ), tal que  $\sum_{i=1}^{60} n_i = N$  é dada por,

$$P([\#\mathbf{Y}_i = n_i]_{i=1}^{60}) = \frac{N!}{(\prod_{i=1}^{60} n_i!)} \prod_{i=1}^{60} p_i^{n_i}. \quad (2.1)$$

As proporções observadas dessa amostra serão então os estimadores de máxima verossimilhança da multinomial, dados por,

$$\hat{p}_i = \frac{n_i}{N}, \quad \forall i = 1, \dots, 60. \quad (2.2)$$

Sabe-se que as posições dentro de cada códon possuem uma dependência entre si, e essa dependência será inserida nos modelos propostos. Para isso, as três posições do

códon serão codificadas inicialmente nesse capítulo pela classificação das bases nitrogenadas como *purina* (A e G) ou *pirimidina* (T ou C). Assim, cada  $Y_{ki}$ , para todo  $k = 1, 2, 3$  e  $i = 1, \dots, 60$  é definido,

$$Y_{ki} = \begin{cases} 1, & \text{purina (A,G);} \\ 0, & \text{pirimidina (T,C).} \end{cases} \quad (2.3)$$

Seja também o vetor  $\mathbf{X}_i = (X_{1i}, X_{2i}, X_{3i})$  das covariáveis (AARISK, AVDIST, TSCORE) do  $i$ -ésimo códon, descritas anteriormente, a probabilidade desse códon é condicionada a elas e denotada por,

$$P(\text{observar o códon } \mathbf{Y}_i) = P(\mathbf{Y}_i | \mathbf{X}_i). \quad (2.4)$$

É importante lembrar que  $\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i) = 1$ , portanto, dividir a probabilidade de observar um códon  $\mathbf{Y}_i$  pela soma das probabilidades dos 60 códons efetivos equivale a dividir a probabilidade desse códon por 1, assim,

$$P(\text{observar o códon } \mathbf{Y}_i) = \frac{P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)} = WP(\mathbf{Y}_i | \mathbf{X}_i). \quad (2.5)$$

Essa probabilidade  $WP(\mathbf{Y}_i | \mathbf{X}_i)$ , ponderada com respeito aos 60 códons efetivos é de extrema importância para realizar as estimativas das probabilidades dos códons, pois garante que a soma delas seja sempre 1. A equação (2.1) é reescrita fazendo uso dessa normalização e considerando as covariáveis,

$$\begin{aligned} P([\#\mathbf{Y}_i = n_i]_{i=1}^{60}) &= \frac{N!}{(\prod_{i=1}^{60} n_i!)} \prod_{i=1}^{60} [WP(\mathbf{Y}_i | \mathbf{X}_i)]^{n_i} \\ &= \frac{N!}{(\prod_{i=1}^{60} n_i!)} \prod_{i=1}^{60} \left[ \frac{P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)} \right]^{n_i}. \end{aligned} \quad (2.6)$$

## 2.1 Modelos Lineares Generalizados

Os modelos lineares generalizados (MLGen) (Nelder e Wedderburn, 1972) são uma extensão dos modelos lineares, cujas respostas seguem uma distribuição normal, para variáveis respostas com outras distribuições, pertencentes à família exponencial. Há

três componentes que especificam um modelo linear generalizado: uma componente aleatória, que identifica a variável resposta  $Y$  e sua distribuição de probabilidade; uma componente sistemática, associada às covariáveis de um preditor linear; uma função de ligação especificando uma função de  $E(Y)$ , que pelo modelo será igualada à componente sistemática.

Para que um conjunto de observações independentes  $(y_1, \dots, y_n)$  possa ser modelado através de um MLGen, é necessário que a distribuição de cada  $y_i$  pertença à família exponencial.

Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma função de distribuição  $f(y_i|\boldsymbol{\theta})$ , tal que  $\boldsymbol{\theta}$  é um vetor de  $k$  parâmetros desconhecidos;  $f(y_i|\boldsymbol{\theta})$  pertence à família exponencial se é possível escrevê-la na seguinte forma (Casella e Berger, 2002):

$$f(y_i|\boldsymbol{\theta}) = h(y_i)c(\boldsymbol{\theta}) \exp \left[ \sum_{j=1}^k w_j(\boldsymbol{\theta})t_j(y_i) \right], \quad (2.7)$$

ou através de uma parametrização diferente para o caso uniparamétrico (Agresti, 2002),

$$f(y_i|\theta) = a(\theta)b(y_i) \exp [y_iQ(\theta)]. \quad (2.8)$$

Nessa parametrização dada pela equação (2.8),  $Q(\theta)$  é chamado de parâmetro natural. A componente sistemática de um MLGen relaciona um vetor  $(\eta_1, \dots, \eta_n)$  às covariáveis, denotadas por  $(x_{i1}, \dots, x_{ip})$ , da seguinte forma:

$$\eta_i = \sum_{l=1}^p \beta_l x_{il}, \quad i = 1, \dots, n. \quad (2.9)$$

Essa combinação linear das covariáveis é chamada de preditor linear. A terceira componente de um MLGen é a função de ligação, que conecta as componentes aleatória e sistemática. Seja  $\mu_i = E(Y_i)$ ,  $i = 1, \dots, n$ , o MLGen conecta  $\mu_i$  à  $\eta_i$  através da função de ligação  $g$ , suposta duplamente diferenciável na seguinte forma,

$$g(\mu_i) = \eta_i \sum_{l=1}^p \beta_l x_{il}, \quad i = 1, \dots, n. \quad (2.10)$$

Quando a função de ligação é identidade, ou seja,  $g(\mu_i) = \mu_i$ , ela especifica um modelo linear para a média, como é feito em modelos de regressão lineares, com respostas normais. Quando a função de ligação transforma a média no parâmetro natural, ou seja,  $g(\mu_i) = Q(\theta)$ , ela é chamada de função de ligação natural ou canônica.

### 2.1.1 Distribuição de Bernoulli e Binomial

É simples de verificar que tanto a distribuição de bernoulli quanto binomial pertencem à família exponencial. Seja uma amostra  $(y_1, \dots, y_n) \stackrel{iid}{\sim} ber(\pi)$ ,

$$\begin{aligned} f(y_i|\pi) &= \pi^{y_i}(1-\pi)^{1-y_i}\mathbb{I}_{\{0,1\}}(y_i) \\ &= (1-\pi)\left[\frac{\pi}{(1-\pi)}\right]^{y_i}\mathbb{I}_{\{0,1\}}(y_i) \\ &= (1-\pi)\exp\left\{y_i\log\left[\frac{\pi}{(1-\pi)}\right]\right\}\mathbb{I}_{\{0,1\}}(y_i), \end{aligned} \quad (2.11)$$

e seja  $w = \sum_{i=1}^n y_i \sim B(n, \pi)$ ,

$$\begin{aligned} f(w|\pi) &= \binom{n}{w}\pi^w(1-\pi)^{n-w}\mathbb{I}_{\{0,1,\dots,n\}}(w) \\ &= \binom{n}{w}(1-\pi)^n\left[\frac{\pi}{(1-\pi)}\right]^w\mathbb{I}_{\{0,1,\dots,n\}}(w) \\ &= \binom{n}{w}(1-\pi)^n\exp\left\{w\log\left[\frac{\pi}{(1-\pi)}\right]\right\}\mathbb{I}_{\{0,1,\dots,n\}}(w). \end{aligned} \quad (2.12)$$

Para ambas as distribuições, o parâmetro natural é dado por  $\log\left[\frac{\pi}{(1-\pi)}\right]$ .

## 2.2 Funções de Ligação mais Comuns em Modelos Binomiais

Retomando as variáveis de interesse deste trabalho, cada posição  $Y_{ki}$  do códon está associada a uma probabilidade  $\pi_{ki} = \pi_k(\mathbf{x}_i)$ , função das covariáveis. Por isso, são definidas as funções de ligação  $\theta_{ki} = g(\pi_{ki})$ . Dessa forma, uma vez estimado  $\theta_{ki}$ , obtém-se  $\hat{\pi}_{ki} = g^{-1}(\hat{\theta}_{ki})$ . As três funções de ligação muito usadas para variáveis binárias são o logito, probito e log-log complementar.

### 2.2.1 Logito

A função logito é a ligação natural dos modelos binomiais, e é definida pela equação:

$$\theta_{ki} = \log \left( \frac{\pi_{ki}}{1 - \pi_{ki}} \right), \quad (2.13)$$

portanto,

$$\pi_{ki} = \frac{e^{\theta_{ki}}}{(1 + e^{\theta_{ki}})}. \quad (2.14)$$

### 2.2.2 Probit

A função de ligação probito é definida pela equação:

$$\theta_{ki} = \Phi^{-1}(\pi_{ki}), \quad (2.15)$$

em que  $\Phi(\cdot)$  é a função de probabilidade acumulada de uma  $N(0, 1)$ , portanto,

$$\pi_{ki} = \Phi(\theta_{ki}). \quad (2.16)$$

### 2.2.3 Log-Log Complementar

A função de ligação log-log complementar é definida pela equação:

$$\theta_{ki} = \log[-\log(1 - \pi_{ki})], \quad (2.17)$$

portanto,

$$\pi_{ki} = 1 - \exp(-e^{\theta_{ki}}). \quad (2.18)$$

É interessante mencionar que as ligações logito e probito assumem que a curva de probabilidade é simétrica com relação às covariáveis, já o log-log complementar assume assimetria dessa curva. Também é importante lembrar, para interpretações dos parâmetros mais adiante, que para as três funções de ligação, as probabilidades  $\pi_{ki}$  são crescentes em  $\theta_{ki}$ .

## 2.3 Modelos Logísticos Regressivos

Os modelos logísticos regressivos (Bonney, 1986, 1987; Bonney et al., 1989; Bonney et al., 1994) introduzem a dependência entre as posições do modelo fazendo uso das posições anteriores como covariáveis para estimar as funções de ligação. Esses modelos serão apresentados em ordem de complexidade (número de parâmetros).

Fazendo uso do teorema da multiplicação, a probabilidade dos códons é fatorada da seguinte maneira,

$$\begin{aligned} P(\mathbf{Y}_i|\mathbf{X}_i) &= P(Y_{1i}, Y_{2i}, Y_{3i}|\mathbf{X}_i) \\ &= P(Y_{1i}|\mathbf{X}_i)P(Y_{2i}|Y_{1i}, \mathbf{X}_i)P(Y_{3i}|Y_{1i}, Y_{2i}, \mathbf{X}_i). \end{aligned} \quad (2.19)$$

Considerando que a probabilidade de cada posição do códon, condicionada às posições anteriores e às covariáveis é dada pela distribuição de *Bernoulli*,

$$(Y_{1i}|\mathbf{X}_i) \sim \text{ber}(\pi_{1i}); \quad (2.20)$$

$$(Y_{2i}|Y_{1i}, \mathbf{X}_i) \sim \text{ber}(\pi_{2i}); \quad (2.21)$$

$$(Y_{3i}|Y_{1i}, Y_{2i}, \mathbf{X}_i) \sim \text{ber}(\pi_{3i}). \quad (2.22)$$

A probabilidade de cada códon na equação (2.19) é então dada por,

$$P(\mathbf{Y}_i|\mathbf{X}_i) = \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}, \quad (2.23)$$

assim, a equação (2.6) é reescrita como,

$$P([\#\mathbf{Y}_i = n_i]_{i=1}^{60}) = \frac{N!}{(\prod_{i=1}^{60} n_i!)} \prod_{i=1}^{60} \left[ \frac{\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \right]^{n_i}. \quad (2.24)$$

A verossimilhança dos modelos, para uma amostra de tamanho  $N$  (tal que  $\sum_{i=1}^{60} n_i = N$ ) é dada por,

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{60} \left[ \frac{\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \right]^{n_i}, \quad (2.25)$$



e a log-verossimilhança,

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 [y_{ki} \log(\pi_{ki}) + (1 - y_{ki}) \log(1 - \pi_{ki})] \\ &- \sum_{i=1}^{60} n_i \log \left[ \sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right]. \end{aligned} \quad (2.26)$$

Os modelos regressivos são quatro, e a diferença entre eles é a forma com que as posições são consideradas na função de ligação.

### 2.3.1 Modelo Independente

O modelo independente assume que não há estrutura de dependência entre as posições do códon.

$$\begin{aligned} \theta_{1i} &= \alpha_1 + \sum_{p=1}^3 \beta_p X_{pi}; \\ \theta_{2i} &= \alpha_2 + \sum_{p=1}^3 \beta_p X_{pi}; \\ \theta_{3i} &= \alpha_3 + \sum_{p=1}^3 \beta_p X_{pi}. \end{aligned} \quad (2.27)$$

Esse modelo possui um total de 6 parâmetros,  $(\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)$ , e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_k} \ell(\boldsymbol{\theta}) = 0, & \forall k = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3. \end{cases} \quad (2.28)$$

Neste modelo, os parâmetros  $\beta_p$ , para todo  $p = 1, 2, 3$  representam a influência das covariáveis (AARISK, AVDIST e TSCORE) nas probabilidades de cada posição do códon. Quando  $\beta_p$  é positivo, como as três covariáveis assumem valores não negativos, então a covariável tem influência crescente nas probabilidades estimadas. Da mesma maneira, quando  $\beta_p$  é negativo, a covariável tem influência decrescente nas probabilidades estimadas.

### 2.3.2 Modelo Igualmente Preditivo

O modelo igualmente preditivo assume que a influência das posições anteriores na seguinte é a mesma.

$$\begin{aligned}
 \theta_{1i} &= \alpha_1 + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{2i} &= \alpha_2 + \gamma Y_{1i} + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{3i} &= \alpha_3 + \gamma (Y_{1i} + Y_{2i}) + \sum_{p=1}^3 \beta_p X_{pi}.
 \end{aligned} \tag{2.29}$$

Esse modelo possui um total de 7 parâmetros,  $(\alpha_1, \alpha_2, \alpha_3, \gamma, \beta_1, \beta_2, \beta_3)$ , e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_k} \ell(\boldsymbol{\theta}) = 0, & \forall k = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\ \frac{\partial}{\partial \gamma} \ell(\boldsymbol{\theta}) = 0. \end{cases} \tag{2.30}$$

Neste modelo, além da influência das covariáveis, já explicada anteriormente, há a influência das posições anteriores nas probabilidades, representada pelo parâmetro  $\gamma$ . Como as posições do códon  $Y_{ki}$  são indicadoras, ou seja, assumem 1 ou 0, valores positivos para  $\gamma$  indicam influência crescente das posições anteriores nas probabilidades estimadas, e valores negativos de  $\gamma$  indicam influência decrescente das posições anteriores.

### 2.3.3 Estrutura Markoviana de Primeira Ordem

A estrutura markoviana de primeira ordem assume que a dependência entre as posições é apenas sobre a posição imediatamente anterior.

$$\begin{aligned}
 \theta_{1i} &= \alpha_1 + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{2i} &= \alpha_2 + \gamma_1 Y_{1i} + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{3i} &= \alpha_3 + \gamma_2 Y_{2i} + \sum_{p=1}^3 \beta_p X_{pi}.
 \end{aligned} \tag{2.31}$$

Esse modelo possui um total de 8 parâmetros,  $(\alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \beta_1, \beta_2, \beta_3)$ , e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_k} \ell(\boldsymbol{\theta}) = 0, & \forall k = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\ \frac{\partial}{\partial \gamma_s} \ell(\boldsymbol{\theta}) = 0, & \forall s = 1, 2. \end{cases} \tag{2.32}$$

Neste modelo, a interpretação dos parâmetros  $\gamma_1$  e  $\gamma_2$  é igual à feita no modelo igualmente preditivo. A diferença é que aqui os parâmetros levam em consideração apenas a posição imediatamente anterior àquela que está sendo modelada.

### 2.3.4 Modelo Aditivo

O modelo aditivo assume que cada posição do códon depende de todas as posições anteriores.

$$\begin{aligned}
 \theta_{1i} &= \alpha_1 + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{2i} &= \alpha_2 + \gamma_1 Y_{1i} + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{3i} &= \alpha_3 + \gamma_2 Y_{1i} + \gamma_3 Y_{2i} + \sum_{p=1}^3 \beta_p X_{pi}.
 \end{aligned} \tag{2.33}$$

Esse modelo possui um total de 9 parâmetros,  $(\alpha_1, \alpha_2, \alpha_3, \gamma_1, \gamma_2, \gamma_3, \beta_1, \beta_2, \beta_3)$ , e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_k} \ell(\boldsymbol{\theta}) = 0, & \forall k = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\ \frac{\partial}{\partial \gamma_s} \ell(\boldsymbol{\theta}) = 0, & \forall s = 1, 2, 3. \end{cases} \quad (2.34)$$

Neste modelo, a interpretação dos parâmetros  $\gamma_1$ ,  $\gamma_2$  e  $\gamma_3$  é igual à feita no modelo igualmente preditivo e na estrutura markoviana de primeira ordem, porém agora levando em consideração todas as posições anteriores àquela modelada.

### 2.3.5 Gradiente e Informação de Fisher

Seja  $\nu_r$  um dentre os parâmetros do vetor  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$  de parâmetros, cada elemento do vetor gradiente para cada modelo é dado por,

$$\begin{aligned} \frac{\partial}{\partial \nu_r} \ell(\boldsymbol{\theta}) &= \frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} n_i \sum_{k=1}^3 [y_{ki} \log(\pi_{ki}) + (1 - y_{ki}) \log(1 - \pi_{ki})] \\ &- \frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} n_i \log \left[ \sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right]. \end{aligned} \quad (2.35)$$

Como apenas as probabilidades  $\pi_{ki}$  são funções dos parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ ,

$$\begin{aligned}
\frac{\partial}{\partial \nu_r} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ y_{ki} \frac{\partial}{\partial \nu_r} \log(\pi_{ki}) + (1 - y_{ki}) \frac{\partial}{\partial \nu_r} \log(1 - \pi_{ki}) \right] \\
&- \sum_{i=1}^{60} n_i \frac{\partial}{\partial \nu_r} \log \left[ \sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \\
&= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \frac{y_{ki}}{\pi_{ki}} \frac{\partial}{\partial \nu_r} \pi_{ki} - \frac{(1 - y_{ki})}{(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} \right] \\
&- \sum_{i=1}^{60} n_i \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \\
&= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \frac{[y_{ki}(1 - \pi_{ki}) - (1 - y_{ki})\pi_{ki}]}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} \\
&- \sum_{i=1}^{60} n_i \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \\
&= \sum_{i=1}^{60} n_i \left[ \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} - \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \right], \quad (2.36)
\end{aligned}$$

em que

$$\begin{aligned}
\frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i) &= \sum_{k=1}^3 \left[ \frac{\partial}{\partial \nu_r} \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \prod_{s \neq k} \pi_{si}^{y_{si}} (1 - \pi_{si})^{1-y_{si}} \\
&= \sum_{k=1}^3 \left[ \frac{y_{ki}}{\pi_{ki}} \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \frac{\partial}{\partial \nu_r} \pi_{ki} - (1 - y_{ki}) \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{-y_{ki}} \frac{\partial}{\partial \nu_r} \pi_{ki} \right] \\
&\times \prod_{s \neq k} \pi_{si}^{y_{si}} (1 - \pi_{si})^{1-y_{si}} \\
&= \sum_{k=1}^3 \frac{[y_{ki}(1 - \pi_{ki}) - (1 - y_{ki})\pi_{ki}]}{\pi_{ki}(1 - \pi_{ki})} \left( \frac{\partial}{\partial \nu_r} \pi_{ki} \right) \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \\
&= \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki}. \quad (2.37)
\end{aligned}$$

Sejam  $\nu_r$  e  $\nu_l$  parâmetros do vetor  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$  de parâmetros, a matriz de informação de Fisher é dada por,

$$I_F(\boldsymbol{\theta}) = -\mathbb{E} \left( \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) \right), \quad (2.38)$$

tal que

$$\begin{aligned} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) &= \frac{\partial}{\partial \nu_l} \sum_{i=1}^{60} n_i \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} \\ &\quad - \frac{\partial}{\partial \nu_l} \sum_{i=1}^{60} n_i \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)}. \end{aligned} \quad (2.39)$$

Como apenas as probabilidades  $\pi_{ki}$  são funções dos parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ ,

$$\begin{aligned} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left\{ \left[ \frac{\partial}{\partial \nu_l} \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \right] \frac{\partial}{\partial \nu_r} \pi_{ki} + \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \left[ \frac{\partial^2}{\partial \nu_r \partial \nu_l} \pi_{ki} \right] \right\} \\ &\quad - \sum_{i=1}^{60} n_i \frac{\partial}{\partial \nu_l} \left[ \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)} \right] \\ &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \pi_{ki} \\ &\quad - \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left\{ \frac{\pi_{ki}(1 - \pi_{ki}) + (y_{ki} - \pi_{ki})(1 - 2\pi_{ki})}{[\pi_{ki}(1 - \pi_{ki})]^2} \right\} \left( \frac{\partial}{\partial \nu_l} \pi_{ki} \right) \left( \frac{\partial}{\partial \nu_r} \pi_{ki} \right) \\ &\quad - \sum_{i=1}^{60} n_i \left\{ \frac{[\sum_{i=1}^{60} \frac{\partial^2}{\partial \nu_r \partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i)]}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}} \right\} \\ &\quad + \sum_{i=1}^{60} n_i \left\{ \frac{[\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)] [\sum_{i=1}^{60} \frac{\partial}{\partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i)]}{[\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}]^2} \right\}, \end{aligned} \quad (2.40)$$

em que,

$$\begin{aligned} \frac{\partial^2}{\partial \nu_r \partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i) &= \frac{\partial}{\partial \nu_l} \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} \\ &= \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \\ &\quad \times \left[ \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_l} \pi_{ki} \right] \left[ \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial}{\partial \nu_r} \pi_{ki} \right] \\ &\quad - \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \\ &\quad \times \sum_{k=1}^3 \left\{ \frac{\pi_{ki}(1 - \pi_{ki}) + (y_{ki} - \pi_{ki})(1 - 2\pi_{ki})}{[\pi_{ki}(1 - \pi_{ki})]^2} \right\} \left( \frac{\partial}{\partial \nu_l} \pi_{ki} \right) \left( \frac{\partial}{\partial \nu_r} \pi_{ki} \right) \\ &\quad + \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] \sum_{k=1}^3 \frac{(y_{ki} - \pi_{ki})}{\pi_{ki}(1 - \pi_{ki})} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \pi_{ki}. \end{aligned} \quad (2.41)$$

Portanto, como  $\mathbb{E}(y_{ki} - \pi_{ki}) = 0$ , os elementos da matriz de informação de Fisher são,

$$\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2}{\partial\nu_r\partial\nu_l}\ell(\boldsymbol{\theta})\right) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \frac{\left(\frac{\partial}{\partial\nu_r}\pi_{ki}\right)\left(\frac{\partial}{\partial\nu_l}\pi_{ki}\right)}{\pi_{ki}(1-\pi_{ki})} \\
&+ \sum_{i=1}^{60} n_i \mathbb{E}\left(\frac{\sum_{i=1}^{60} \frac{\partial^2}{\partial\nu_r\partial\nu_l} P(\mathbf{Y}_i|\mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1-\pi_{ki})^{1-y_{ki}}}\right) \\
&- \sum_{i=1}^{60} n_i \mathbb{E}\left(\frac{\left[\sum_{i=1}^{60} \frac{\partial}{\partial\nu_r} P(\mathbf{Y}_i|\mathbf{X}_i)\right] \left[\sum_{i=1}^{60} \frac{\partial}{\partial\nu_l} P(\mathbf{Y}_i|\mathbf{X}_i)\right]}{\left[\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1-\pi_{ki})^{1-y_{ki}}\right]^2}\right). \quad (2.42)
\end{aligned}$$

As primeiras e segundas derivadas de  $\pi_{ki}$  com relação a  $\nu_r$  e  $\nu_l$  são diferentes para cada uma das funções de ligação explicadas na Seção 2.1. A seguir estão os resultados para cada uma das funções de ligação consideradas.

**Logito:**

$$\frac{\partial}{\partial\nu_r}\pi_{ki} = \pi_{ki}(1-\pi_{ki})\frac{\partial}{\partial\nu_r}\theta_{ki}; \quad (2.43)$$

$$\frac{\partial^2}{\partial\nu_r\partial\nu_l}\pi_{ki} = \pi_{ki}(1-\pi_{ki})(1-2\pi_{ki})\left(\frac{\partial}{\partial\nu_r}\theta_{ki}\right)\left(\frac{\partial}{\partial\nu_l}\theta_{ki}\right). \quad (2.44)$$

**Probit:**

$$\frac{\partial}{\partial\nu_r}\pi_{ki} = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\theta_{ki}^2}{2}\right) \frac{\partial}{\partial\nu_r}\theta_{ki}; \quad (2.45)$$

$$\frac{\partial^2}{\partial\nu_r\partial\nu_l}\pi_{ki} = \frac{-1}{\sqrt{2\pi}} \exp\left(\frac{-\theta_{ki}^2}{2}\right) \theta_{ki} \left(\frac{\partial}{\partial\nu_r}\theta_{ki}\right) \left(\frac{\partial}{\partial\nu_l}\theta_{ki}\right). \quad (2.46)$$

**Log-Log Complementar:**

$$\frac{\partial}{\partial\nu_r}\pi_{ki} = \exp(\theta_{ki} - e^{\theta_{ki}}) \frac{\partial}{\partial\nu_r}\theta_{ki}; \quad (2.47)$$

$$\frac{\partial^2}{\partial\nu_r\partial\nu_l}\pi_{ki} = (1 - e^{\theta_{ki}}) \left(\frac{\partial}{\partial\nu_r}\theta_{ki}\right) \left(\frac{\partial}{\partial\nu_l}\theta_{ki}\right) \exp(\theta_{ki} - e^{\theta_{ki}}). \quad (2.48)$$

## 2.4 Modelo Baseado na Representação de Bahadur

O modelo baseado na representação de Bahadur (Bahadur, 1961) considera a dependência entre as posições dentro de um códon de maneira diferente dos modelos

regressivos. Esse modelo não faz uso das posições anteriores como covariáveis para estimar a função de ligação, mas estima as probabilidades de ser uma *purina* ou *pirimidina* em cada posição e as correlações entre as posições dentro do códon. É claro que não é possível calcular diretamente a correlação entre variáveis binárias como têm-se aqui, e para isso uma normalização é feita sobre essas variáveis, como será explicado a seguir.

Assumindo independência entre as posições do códon, a probabilidade pode ser escrita exatamente igual à equação (2.23), em que cada  $\pi_{ki} = g^{-1}(\theta_{ki})$  é obtido pelo modelo independente, descrito na equação (2.28), para qualquer uma das três funções de ligação, ou seja,  $\theta_{ki} = g(\pi_{ki}) = \alpha_k + \sum_{p=1}^3 \beta_p X_{pi}$ ,

$$P_I(\mathbf{Y}_i | \mathbf{X}_i) = \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}. \quad (2.49)$$

Para utilizar a representação de Bahadur, define-se a seguinte variável de normalização:

$$U_{ki} = \frac{Y_{ki} - \pi_{ki}}{\sqrt{\pi_{ki}(1 - \pi_{ki})}}, \quad (2.50)$$

de forma que  $\rho_{12} = E(U_{1i}U_{2i})$ ,  $\rho_{13} = E(U_{1i}U_{3i})$  e  $\rho_{23} = E(U_{2i}U_{3i})$  são as correlações entre as posições dentro do códon, e  $\rho_{123} = E(U_{1i}U_{2i}U_{3i})$  uma medida de correlação que relaciona as três posições simultaneamente.

Finalmente, a representação de Bahadur introduz a estrutura de dependência no modelo da seguinte forma:

$$\begin{aligned} P_B(\mathbf{Y}_i | \mathbf{X}_i) &= P_I(\mathbf{Y}_i | \mathbf{X}_i) f(\boldsymbol{\rho}, \mathbf{u}_i) \\ &= \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] f(\boldsymbol{\rho}, \mathbf{u}_i), \end{aligned} \quad (2.51)$$

em que,

$$f(\boldsymbol{\rho}, \mathbf{u}_i) = 1 + \rho_{12}u_{1i}u_{2i} + \rho_{13}u_{1i}u_{3i} + \rho_{23}u_{2i}u_{3i} + \rho_{123}u_{1i}u_{2i}u_{3i}. \quad (2.52)$$

Para garantir que  $P_B(\mathbf{Y}_i | \mathbf{X}_i)$  defina uma medida de probabilidade, e sabendo-se que  $P_I(\mathbf{Y}_i | \mathbf{X}_i)$  é estritamente positiva, a restrição a seguir deve ser obedecida.



$$\text{R1. } f(\boldsymbol{\rho}, \mathbf{U}_i) > 0, \quad \forall i = 1, \dots, 60.$$

Portanto, no modelo baseado na representação de Bahadur, o interesse é estimar o conjunto  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \rho_{12}, \rho_{13}, \rho_{23}, \rho_{123})$  de parâmetros, e uma segunda restrição deve ser obedecida, uma vez que  $\rho_{12}$ ,  $\rho_{13}$ ,  $\rho_{23}$  e  $\rho_{123}$  definem correlações.

$$\text{R2. } \rho_{12}, \rho_{13}, \rho_{23}, \rho_{123} \in [-1, 1].$$

Assim, utilizando  $P_B(\mathbf{Y}_i|\mathbf{X}_i)$  na equação (2.6),

$$P([\#\mathbf{Y}_i = n_i | \mathbf{X}_i]_{i=1}^{60}) = \frac{N!}{(\prod_{i=1}^{60} n_i!)} \times \prod_{i=1}^{60} \left\{ \frac{[\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}] f(\boldsymbol{\rho}, \mathbf{u}_i)}{\sum_{i=1}^{60} [\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}] f(\boldsymbol{\rho}, \mathbf{u}_i)} \right\}^{n_i} \quad (2.53)$$

Portanto a verossimilhança será

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) \propto \prod_{i=1}^{60} \left\{ \frac{[\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}] f(\boldsymbol{\rho}, \mathbf{u}_i)}{\sum_{i=1}^{60} [\prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}}] f(\boldsymbol{\rho}, \mathbf{u}_i)} \right\}^{n_i}, \quad (2.54)$$

e a log-verossimilhança,

$$\begin{aligned} \ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 [y_{ki} \log(\pi_{ki}) + (1 - y_{ki}) \log(1 - \pi_{ki})] \\ &+ \sum_{i=1}^{60} n_i \log[f(\boldsymbol{\rho}, \mathbf{u}_i)] \\ &- \sum_{i=1}^{60} n_i \log \left\{ \sum_{i=1}^{60} \left[ \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \right] f(\boldsymbol{\rho}, \mathbf{u}_i) \right\}. \end{aligned} \quad (2.55)$$

Finalmente, ao maximizar  $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$  sob as restrições R1 e R2, obtém-se os estimadores de máxima verossimilhança para os parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$ .

Como a maximização da log-verossimilhança sob as duas restrições para estimar os parâmetros simultaneamente é um problema computacionalmente complexo, a estimação é feita em dois passos. No primeiro passo estima-se  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  através do modelo

independente descrito na seção de modelos regressivos, para alguma das três funções de ligação.

Em seguida, utilizando  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$  estimados no passo 1, obtém-se  $\hat{\pi}_{ki}$  para todo  $k = 1, 2, 3$  e  $i = 1, \dots, 60$ . A partir dessas probabilidades, calculam-se as probabilidades dos códons assumindo independência entre as posições,  $\hat{P}_I(\mathbf{Y}_i | \mathbf{X}_i)$ .

As variáveis para a estrutura de dependência também são calculadas a partir das estimativas  $\hat{\pi}_{ki}$ , i.e.,

$$\hat{U}_{ki} = \frac{Y_{ki} - \hat{\pi}_{ki}}{\sqrt{\hat{\pi}_{ki}(1 - \hat{\pi}_{ki})}}. \quad (2.56)$$

Finalmente, no segundo passo estima-se  $\boldsymbol{\rho} | (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ . Como  $\pi_{ki}$  já foram estimados, o primeiro termo da equação (2.55) é constante com relação a  $\boldsymbol{\rho}$ , portanto a log-verossimilhança a ser maximizada é a seguinte,

$$\ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left\{ \log [f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)] - \log \left[ \sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right] \right\}. \quad (2.57)$$

Assim, para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir, sob as restrições R1 e R2.

$$\begin{cases} \frac{\partial}{\partial \rho_{12}} \ell(\boldsymbol{\rho}) = 0; \\ \frac{\partial}{\partial \rho_{13}} \ell(\boldsymbol{\rho}) = 0; \\ \frac{\partial}{\partial \rho_{23}} \ell(\boldsymbol{\rho}) = 0; \\ \frac{\partial}{\partial \rho_{123}} \ell(\boldsymbol{\rho}) = 0. \end{cases} \quad (2.58)$$

### 2.4.1 Gradiente e Informação de Fisher

Os elementos do vetor gradiente de cada modelo são dados por,

$$\frac{\partial}{\partial \rho_{12}} \ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left[ \frac{\hat{u}_{1i} \hat{u}_{2i}}{f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} - \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} \hat{u}_{1i} \hat{u}_{2i}}{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \right]; \quad (2.59)$$

$$\frac{\partial}{\partial \rho_{13}} \ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left[ \frac{\hat{u}_{1i} \hat{u}_{3i}}{f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} - \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} \hat{u}_{1i} \hat{u}_{3i}}{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \right]; \quad (2.60)$$

$$\frac{\partial}{\partial \rho_{23}} \ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left[ \frac{\hat{u}_{2i} \hat{u}_{3i}}{f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} - \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} \hat{u}_{2i} \hat{u}_{3i}}{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \right]; \quad (2.61)$$

$$\frac{\partial}{\partial \rho_{123}} \ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left[ \frac{\hat{u}_{1i} \hat{u}_{2i} \hat{u}_{3i}}{f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} - \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} \hat{u}_{1i} \hat{u}_{2i} \hat{u}_{3i}}{\sum_{i=1}^{60} \prod_{k=1}^3 \hat{\pi}_{ki}^{y_{ki}} (1 - \hat{\pi}_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \right]. \quad (2.62)$$

A matriz de informação de Fisher é dada por,

$$I_F(\boldsymbol{\rho}) = -\mathbb{E} \begin{bmatrix} \frac{\partial^2}{\partial \rho_{12}^2} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{12} \partial \rho_{13}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{12} \partial \rho_{23}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{12} \partial \rho_{123}} \ell(\boldsymbol{\rho}) \\ \frac{\partial^2}{\partial \rho_{12} \partial \rho_{13}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{13}^2} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{13} \partial \rho_{23}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{13} \partial \rho_{123}} \ell(\boldsymbol{\rho}) \\ \frac{\partial^2}{\partial \rho_{12} \partial \rho_{23}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{13} \partial \rho_{23}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{23}^2} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{23} \partial \rho_{123}} \ell(\boldsymbol{\rho}) \\ \frac{\partial^2}{\partial \rho_{12} \partial \rho_{123}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{13} \partial \rho_{123}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{23} \partial \rho_{123}} \ell(\boldsymbol{\rho}) & \frac{\partial^2}{\partial \rho_{123}^2} \ell(\boldsymbol{\rho}) \end{bmatrix}, \quad (2.63)$$

em que, para  $\rho_{*1}$  e  $\rho_{*2}$  representando quaisquer um dentre os parâmetros de correlação do vetor  $\boldsymbol{\rho}$ ,

$$\begin{aligned} \frac{\partial^2}{\partial \rho_{*1} \partial \rho_{*2}} \ell(\boldsymbol{\rho}) &= - \sum_{i=1}^{60} n_i \frac{\left[ \frac{\partial}{\partial \rho_{*1}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right] \left[ \frac{\partial}{\partial \rho_{*2}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right]}{[f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)]^2} \\ &+ \sum_{i=1}^{60} n_i \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \frac{\partial}{\partial \rho_{*1}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \\ &\times \frac{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} \frac{\partial}{\partial \rho_{*2}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \pi_{ki}^{y_{ki}} (1 - \pi_{ki})^{1-y_{ki}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)}. \end{aligned} \quad (2.64)$$

# Capítulo 3

## Modelos Multinomiais Multivariados

Assim como nos modelos binomiais multivariados discutidos no capítulo anterior, os modelos multinomiais multivariados também levam em consideração a estrutura de dependência entre as posições do códon. A diferença agora é que cada posição  $Y_{ki}$  ( $k = 1, 2, 3$ ) do códon será considerada não mais como *purina* ou *pirimidina*, mas como a base nitrogenada que assume, ou seja, T, C, A ou G.

Codificando cada base numericamente, a relação será  $T = 0$ ,  $C = 1$ ,  $A = 2$  e  $G = 3$ ,  $Y_{ki}$  assume portanto um dentre os valores 0, 1, 2 ou 3. Essa codificação numérica será utilizada somente para facilitar a notação ao longo da explicação e desenvolvimento dos modelos multinomiais, pois as bases nitrogenadas não são valores ordinais e não seria correto definir modelos que considerem as bases numericamente.

Para que as bases nitrogenadas assumidas por  $Y_{ki}$  sejam corretamente consideradas como categorias, as seguintes variáveis indicadoras são criadas:

$$Z_{kji} = \begin{cases} 1, & Y_{ki} = j \\ 0, & Y_{ki} \neq j \end{cases} ; \quad \forall j = 1, 2, 3. \quad (3.1)$$

Assim, a relação de correspondência entre as variáveis  $Y_{ki}$  que assumem a base nitrogenada de cada posição do códon e o vetor  $\mathbf{Z}_{ki} = (Z_{k1i}, Z_{k2i}, Z_{k3i})$  de variáveis

indicadoras é dada pela Tabela 3.1.

Tabela 3.1: Codificação das Bases Nitrogenadas

Nucleotídeo	Codificação	$(Z_{k1i}, Z_{k2i}, Z_{k3i})$
T	0	$(0, 0, 0)$
C	1	$(1, 0, 0)$
A	2	$(0, 1, 0)$
G	3	$(0, 0, 1)$

Os modelos multinomiais consideram as mesmas covariáveis  $\mathbf{X}_i$  já usadas nos modelos binomiais, e as probabilidades a serem estimadas também fazem uso da normalização  $WP(\mathbf{Y}_i|\mathbf{X}_i)$  previamente explicada, e descrita pela equação (2.6).

Outro ponto importante a ser ressaltado é que os modelos multinomiais fazem uso apenas da função de ligação logito, por essa função já ter uso comum em modelos com respostas múltiplas.

$$\text{logito}(\pi_{kji}) = \log\left(\frac{\pi_{kji}}{\pi_{k0i}}\right) = \theta_{kji}. \quad (3.2)$$

Portanto,

$$\left(\frac{\pi_{kji}}{\pi_{k0i}}\right) = e^{\theta_{kji}} \Rightarrow \pi_{kji} = \pi_{k0i}e^{\theta_{kji}}. \quad (3.3)$$

Como  $\pi_{k0i} = 1 - \sum_{j=1}^3 \pi_{kji}$ ,

$$\pi_{k0i} = 1 - \sum_{j=1}^3 \pi_{k0i}e^{\theta_{kji}} \Rightarrow \pi_{k0i} + \pi_{k0i} \sum_{j=1}^3 e^{\theta_{kji}} = 1 \Rightarrow \pi_{k0i} = \frac{1}{1 + \sum_{j=1}^3 e^{\theta_{kji}}}. \quad (3.4)$$

Finalmente,

$$\pi_{kji} = \frac{e^{\theta_{kji}}}{1 + \sum_{j=1}^3 e^{\theta_{kji}}}. \quad (3.5)$$

## 3.1 Modelos Logísticos Regressivos

Fazendo novamente uso do teorema da multiplicação,

$$P(\mathbf{Y}_i|\mathbf{X}_i) = P(Y_{1i}|\mathbf{X}_i)P(Y_{2i}|Y_{1i}, \mathbf{X}_i)P(Y_{3i}|Y_{1i}, Y_{2i}, \mathbf{X}_i).$$

Agora  $(Y_{ki}|Y_{1i}, \dots, Y_{k-1,i}, \mathbf{X}_i) \sim \text{Multinomial}(1, \pi_{k0i}, \pi_{k1i}, \pi_{k2i}, \pi_{k3i})$ , e portanto, fazendo uso da relação entre  $Y_{ki}$  e  $\mathbf{Z}_{ki}$ , a probabilidade de cada códon é dada por,

$$\begin{aligned} P(\mathbf{Y}_i|\mathbf{X}_i) &= \prod_{k=1}^3 \pi_{k0i}^{1-\sum_{j=1}^3 z_{kji}} \pi_{k1i}^{z_{k1i}} \pi_{k2i}^{z_{k2i}} \pi_{k3i}^{z_{k3i}} \\ &= \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})}, \end{aligned} \quad (3.6)$$

e a equação (2.6) pode ser reescrita como:

$$\begin{aligned} P([\#\mathbf{Y}_i = n_i]_{i=1}^{60}) &= \frac{N!}{(\prod_{i=1}^{60} n_i!)} \\ &\times \prod_{i=1}^{60} \left\{ \frac{\prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\}^{n_i}. \end{aligned} \quad (3.7)$$

Finalmente, a verossimilhança é dada por,

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{60} \left\{ \frac{\prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\}^{n_i}, \quad (3.8)$$

e a log-verossimilhança,

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \theta_{kji} - \log(1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right] \\ &- \sum_{i=1}^{60} n_i \log \left[ \sum_{i=1}^{60} \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right]. \end{aligned} \quad (3.9)$$

Os modelos propostos por [Bonney et al. \(1994\)](#) tratavam de fato das variáveis resposta  $Y_{ki}$  como multinomiais, e portanto, os quatro modelos anteriormente descritos para respostas binomiais serão explicados agora para os modelos multinomiais multivariados. Novamente, esses modelos utilizam as posições anteriores como covariáveis para estimar os logitos, e serão apresentados em ordem de complexidade (número de parâmetros).

### 3.1.1 Modelo Independente

Como há quatro categorias de respostas, há portanto 3 logitos. O modelo independente assume que não há estrutura de dependência entre as posições do códon.

$$\begin{aligned}\theta_{1ji} &= \alpha_{1j} + \sum_{p=1}^3 \beta_p X_{pi}; \\ \theta_{2ji} &= \alpha_{2j} + \sum_{p=1}^3 \beta_p X_{pi}; \\ \theta_{3ji} &= \alpha_{3j} + \sum_{p=1}^3 \beta_p X_{pi},\end{aligned}\tag{3.10}$$

para todo  $j = 1, 2, 3$  e  $i = 1, \dots, 60$ . Esse modelo possui um total de 12 parâmetros, e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_{kj}} \ell(\boldsymbol{\theta}) = 0, & \forall k, j = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3. \end{cases}\tag{3.11}$$

Neste modelo, os parâmetros  $\beta_p$ , para todo  $p = 1, 2, 3$  representam a influência das covariáveis (AARISK, AVDIST e TSCORE) nas probabilidades de cada posição do códon. Quando  $\beta_p$  é positivo, como as três covariáveis assumem valores não negativos, então a covariável tem influência crescente nas probabilidades estimadas. Da mesma maneira, quando  $\beta_p$  é negativo, a covariável tem influência decrescente nas probabilidades estimadas.

### 3.1.2 Modelo Igualmente Preditivo

O modelo igualmente preditivo assume que a influência das posições anteriores na seguinte é a mesma. Logo,

$$\begin{aligned}
 \theta_{1ji} &= \alpha_{1j} + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{2ji} &= \alpha_{2j} + \sum_{s=1}^3 \gamma_s Z_{1si} + \sum_{p=1}^3 \beta_p X_{pi}; \\
 \theta_{3ji} &= \alpha_{3j} + \sum_{s=1}^3 \gamma_s \sum_{k=1}^2 Z_{ksi} + \sum_{p=1}^3 \beta_p X_{pi},
 \end{aligned} \tag{3.12}$$

para todo  $j = 1, 2, 3$  e  $i = 1, \dots, 60$ . Esse modelo possui um total de 15 parâmetros, e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases}
 \frac{\partial}{\partial \alpha_{kj}} \ell(\boldsymbol{\theta}) = 0, & \forall k, j = 1, 2, 3; \\
 \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\
 \frac{\partial}{\partial \gamma_s} \ell(\boldsymbol{\theta}) = 0, & \forall s = 1, 2, 3.
 \end{cases} \tag{3.13}$$

Neste modelo, além da influência das covariáveis, já explicada anteriormente, há a influência das posições anteriores nas probabilidades, representada pelos parâmetros  $\gamma_s$ , para todo  $s = 1, 2, 3$ . Como as posições do códon  $Z_{kji}$  são indicadoras, ou seja, assumem 1 ou 0, valores positivos para  $\gamma_s$  indicam influência crescente das posições anteriores nas probabilidades estimadas, e valores negativos de  $\gamma_s$  indicam influência decrescente das posições anteriores.



### 3.1.3 Estrutura Markoviana de Primeira Ordem

A estrutura markoviana de primeira ordem assume que a dependência entre as posições é apenas sobre a posição imediatamente anterior.

$$\begin{aligned}
\theta_{1ji} &= \alpha_{1j} + \sum_{p=1}^3 \beta_p X_{pi}; \\
\theta_{2ji} &= \alpha_{2j} + \sum_{s=1}^3 \gamma_{1s} Z_{1si} + \sum_{p=1}^3 \beta_p X_{pi}; \\
\theta_{3ji} &= \alpha_{3j} + \sum_{s=1}^3 \gamma_{2s} Z_{2si} + \sum_{p=1}^3 \beta_p X_{pi},
\end{aligned} \tag{3.14}$$

para todo  $j = 1, 2, 3$  e  $i = 1, \dots, 60$ . Esse modelo possui um total de 18 parâmetros, e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_{kj}} \ell(\boldsymbol{\theta}) = 0, & \forall k, j = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\ \frac{\partial}{\partial \gamma_{ks}} \ell(\boldsymbol{\theta}) = 0, & \forall k = 1, 2; s = 1, 2, 3. \end{cases} \tag{3.15}$$

Neste modelo, a interpretação dos parâmetros  $\gamma_{1s}$  e  $\gamma_{2s}$ , para todo  $s = 1, 2, 3$ , é igual à feita no modelo igualmente preditivo. A diferença é que aqui os parâmetros levam em consideração apenas a posição imediatamente anterior àquela que está sendo modelada.

### 3.1.4 Modelo Aditivo

O modelo aditivo assume que cada posição do códon depende de todas as posições anteriores.

$$\begin{aligned}
\theta_{1ji} &= \alpha_{1j} + \sum_{p=1}^3 \beta_p X_{pi}; \\
\theta_{2ji} &= \alpha_{2j} + \sum_{s=1}^3 \gamma_{1s} Z_{1si} + \sum_{p=1}^3 \beta_p X_{pi}; \\
\theta_{3ji} &= \alpha_{3j} + \sum_{s=1}^3 (\gamma_{2s} Z_{1si} + \gamma_{3s} Z_{2si}) + \sum_{p=1}^3 \beta_p X_{pi},
\end{aligned} \tag{3.16}$$

para todo  $j = 1, 2, 3$  e  $i = 1, \dots, 60$ . Esse modelo possui um total de 21 parâmetros, e para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\begin{cases} \frac{\partial}{\partial \alpha_{kj}} \ell(\boldsymbol{\theta}) = 0, & \forall k, j = 1, 2, 3; \\ \frac{\partial}{\partial \beta_p} \ell(\boldsymbol{\theta}) = 0, & \forall p = 1, 2, 3; \\ \frac{\partial}{\partial \gamma_{ks}} \ell(\boldsymbol{\theta}) = 0, & \forall k, s = 1, 2, 3. \end{cases} \quad (3.17)$$

Neste modelo, a interpretação dos parâmetros  $\gamma_{1s}$ ,  $\gamma_{2s}$  e  $\gamma_{3s}$ , para todo  $s = 1, 2, 3$ , é igual à feita no modelo igualmente preditivo e na estrutura markoviana de primeira ordem, porém agora levando em consideração todas as posições anteriores àquela modelada.

### 3.1.5 Gradiente e Informação de Fisher

Seja  $\nu_r$  dentre os parâmetros do vetor  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$  de parâmetros, cada elemento do vetor gradiente para cada modelo é dado por,

$$\begin{aligned} \frac{\partial}{\partial \nu_r} \ell(\boldsymbol{\theta}) &= \frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \theta_{kji} - \log \left( 1 + \sum_{j=1}^3 e^{\theta_{kji}} \right) \right] \\ &- \frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} n_i \log \left[ \sum_{i=1}^{30} \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right]. \end{aligned} \quad (3.18)$$

Como apenas os logitos  $\theta_{kji}$  são funções dos parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ ,

$$\begin{aligned}
\frac{\partial}{\partial \nu_r} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \frac{\partial}{\partial \nu_r} \theta_{kji} - \frac{\frac{\partial}{\partial \nu_r} (1 + \sum_{j=1}^3 e^{\theta_{kji}})}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&- \sum_{i=1}^{60} n_i \left\{ \frac{\frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\} \\
&= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \frac{\partial}{\partial \nu_r} \theta_{kji} - \frac{\sum_{j=1}^3 \frac{\partial}{\partial \nu_r} e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&- \sum_{i=1}^{60} n_i \left\{ \frac{\frac{\partial}{\partial \nu_r} \sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\} \\
&= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \\
&- \sum_{i=1}^{60} n_i \left\{ \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\}, \quad (3.19)
\end{aligned}$$

em que

$$\begin{aligned}
\frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i) &= \sum_{k=1}^3 \left[ \frac{\partial}{\partial \nu_r} \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \prod_{s \neq k} \frac{e^{\sum_{j=1}^3 z_{sji} \theta_{sji}}}{(1 + \sum_{j=1}^3 e^{\theta_{sji}})} \\
&= \sum_{k=1}^3 \left[ \frac{(\frac{\partial}{\partial \nu_r} e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) (1 + \sum_{j=1}^3 e^{\theta_{kji}}) - e^{\sum_{j=1}^3 z_{kji} \theta_{kji}} \sum_{j=1}^3 \frac{\partial}{\partial \nu_r} e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})^2} \right] \\
&\times \prod_{s \neq k} \frac{e^{\sum_{j=1}^3 z_{sji} \theta_{sji}}}{(1 + \sum_{j=1}^3 e^{\theta_{sji}})} \\
&= \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \frac{\partial}{\partial \nu_r} \theta_{kji} - \sum_{j=1}^3 \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \frac{\partial}{\partial \nu_r} \theta_{kji} \right] \\
&\times \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \\
&= \left[ \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji}. \quad (3.20)
\end{aligned}$$

Sejam  $\nu_r$  e  $\nu_l$  parâmetros do vetor  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$  de parâmetros, a matriz de informação de Fisher é dada por,

$$I_F(\boldsymbol{\theta}) = -\mathbb{E} \left( \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) \right), \quad (3.21)$$

tal que

$$\begin{aligned} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) &= \frac{\partial}{\partial \nu_l} \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \\ &- \frac{\partial}{\partial \nu_l} \sum_{i=1}^{60} n_i \left\{ \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right\}. \end{aligned} \quad (3.22)$$

Como apenas os logitos  $\theta_{ki}$  são funções dos parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ ,

$$\begin{aligned} \frac{\partial^2}{\partial \nu_r \partial \nu_l} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left\{ \left[ -\frac{\partial}{\partial \nu_l} \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \right\} \\ &+ \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left\{ \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial^2}{\partial \nu_r \partial \nu_l} \theta_{kji} \right\} \\ &- \sum_{i=1}^{60} n_i \frac{\partial}{\partial \nu_l} \left[ \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P(\mathbf{Y}_i | \mathbf{X}_i)} \right] \\ &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial^2}{\partial \nu_r \partial \nu_l} \theta_{kji} \\ &- \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \left[ \frac{\partial}{\partial \nu_l} \theta_{kji} - \frac{\sum_{j=1}^3 e^{\theta_{kji}} \frac{\partial}{\partial \nu_l} \theta_{kji}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{e^{\theta_{kji}} \left( \frac{\partial}{\partial \nu_l} \theta_{kji} \right)}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \\ &- \sum_{i=1}^{60} n_i \frac{\sum_{i=1}^{60} \frac{\partial^2}{\partial \nu_r \partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \\ &+ \sum_{i=1}^{60} n_i \frac{\left[ \sum_{i=1}^{60} \frac{\partial}{\partial \nu_r} P(\mathbf{Y}_i | \mathbf{X}_i) \right] \left[ \sum_{i=1}^{60} \frac{\partial}{\partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i) \right]}{\left\{ \sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right] \right\}^2}, \end{aligned} \quad (3.23)$$

em que,

$$\begin{aligned}
\frac{\partial^2}{\partial \nu_r \partial \nu_l} P(\mathbf{Y}_i | \mathbf{X}_i) &= \left[ \frac{\partial}{\partial \nu_l} \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \\
&+ \left[ \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \frac{\partial}{\partial \nu_l} \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \\
&= \left[ \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_l} \theta_{kji} \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial}{\partial \nu_r} \theta_{kji} \\
&+ \left[ \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{\partial^2}{\partial \nu_r \partial \nu_l} \theta_{kji} \\
&- \left[ \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&\times \sum_{k=1}^3 \sum_{j=1}^3 \left[ \frac{\partial}{\partial \nu_l} \theta_{kji} - \frac{\sum_{j=1}^3 e^{\theta_{kji}} \frac{\partial}{\partial \nu_l} \theta_{kji}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \frac{e^{\theta_{kji}} \left( \frac{\partial}{\partial \nu_l} \theta_{kji} \right)}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})}. \quad (3.24)
\end{aligned}$$

Portanto, como  $\mathbb{E}(z_{kji} - \pi_{kji}) = \mathbb{E} \left( z_{kji} - \frac{e^{\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right) = 0$ , os elementos da matriz

de informação de Fisher são,

$$\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2}{\partial\nu_r\partial\nu_l}\ell(\boldsymbol{\theta})\right) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \sum_{j=1}^3 \frac{e^{\theta_{kji}} \left(\frac{\partial}{\partial\nu_r}\theta_{kji}\right)}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \left[ \frac{\partial}{\partial\nu_l}\theta_{kji} - \frac{\sum_{j=1}^3 e^{\theta_{kji}} \frac{\partial}{\partial\nu_l}\theta_{kji}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] \\
&- \sum_{i=1}^{60} n_i \mathbb{E} \left( \frac{\left[ \sum_{i=1}^{60} \frac{\partial}{\partial\nu_r} P(\mathbf{Y}_i|\mathbf{X}_i) \right] \left[ \sum_{i=1}^{60} \frac{\partial}{\partial\nu_l} P(\mathbf{Y}_i|\mathbf{X}_i) \right]}{\left\{ \sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji}\theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right] \right\}^2} \right) \\
&+ \sum_{i=1}^{60} n_i \mathbb{E} \left( \frac{\sum_{i=1}^{60} \frac{\partial^2}{\partial\nu_r\partial\nu_l} P(\mathbf{Y}_i|\mathbf{X}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji}\theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right]} \right) \quad (3.25)
\end{aligned}$$

## 3.2 Modelo Baseado na Representação de Bahadur

A extensão do modelo baseado na representação de Bahadur para o caso multinomial descreve a probabilidade de cada posição do códon assumir uma dentre as bases T, C, A ou G e a correlação entre as posições e as bases assumidas. Quatro diferentes extensões são apresentadas em ordem de complexidade (número de parâmetros), levando em consideração diferentes maneiras de estudar as correlações no caso de respostas multinomiais.

De maneira análoga ao caso binário, assumindo independência entre as posições do códon, a probabilidade pode ser escrita exatamente igual à equação (3.6) em que cada  $\pi_{kji} = g^{-1}(\theta_{kji})$  é obtido pelo modelo independente descrito na equação (3.10), ou seja,  $\theta_{kji} = g(\pi_{kji}) = \alpha_{kj} + \sum_{p=1}^3 \beta_p X_{pi}$ ,

$$P_I(\mathbf{Y}_i|\mathbf{X}_i) = \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji}\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})}. \quad (3.26)$$

A variável de normalização é definida para todas as posições e categorias,

$$U_{kji} = \frac{Z_{kji} - \pi_{kji}}{\sqrt{\pi_{kji}(1 - \pi_{kji})}}. \quad (3.27)$$

Finalmente, a expansão para a representação de Bahadur no caso de respostas

politômicas será:

$$\begin{aligned} P_B(\mathbf{Y}_i|\mathbf{X}_i) &= P_I(\mathbf{Y}_i|\mathbf{X}_i)f(\boldsymbol{\rho}, \mathbf{u}_i) \\ &= \left[ \prod_{k=1}^3 \frac{\prod_{j=1}^3 e^{z_{kji}\theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} \right] f(\boldsymbol{\rho}, \mathbf{u}_i). \end{aligned} \quad (3.28)$$

Como mencionado anteriormente, quatro extensões do modelo baseado na representação de Bahadur para o caso de respostas politômicas são propostas, e essas extensões são diferenciadas pela estrutura de dependência  $f(\boldsymbol{\rho}, \mathbf{u}_i)$ . Antes de discutir detalhadamente essas estruturas, algumas considerações a respeito do modelo serão feitas.

Assim como nos modelos para respostas binárias, por  $P_B(\mathbf{Y}_i|\mathbf{X}_i)$  definir uma medida de probabilidade, é necessário que esta seja estritamente positiva,

$$\text{R1. } f(\boldsymbol{\rho}, \mathbf{U}_i) > 0, \quad \forall i = 1, \dots, 60.$$

Portanto, no modelo baseado na representação de Bahadur, o interesse é estimar o conjunto  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$  de parâmetros, e uma segunda restrição deve ser obedecida, uma vez que  $\boldsymbol{\rho}$  definem correlações,

$$\text{R2. } \rho_* \in [-1, 1], \quad \forall \rho_* \in \boldsymbol{\rho}.$$

Assim, utilizando  $P_B(\mathbf{Y}_i|\mathbf{X}_i)$  na equação (2.6), a verossimilhança é dada por,

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) \propto \prod_{i=1}^{60} \left\{ \frac{\prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji}\theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right] f(\boldsymbol{\rho}, \mathbf{u}_i)}{\sum_{i=1}^{60} \prod_{k=1}^3 \left[ (e^{\sum_{j=1}^3 z_{kji}\theta_{kji}}) / (1 + \sum_{j=1}^3 e^{\theta_{kji}}) \right] f(\boldsymbol{\rho}, \mathbf{u}_i)} \right\}^{n_i} \quad (3.29)$$

e a log-verossimilhança,

$$\begin{aligned}
\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}) &= \sum_{i=1}^{60} n_i \sum_{k=1}^3 \left[ \sum_{j=1}^3 z_{kji} \theta_{kji} - \log \left( 1 + \sum_{j=1}^3 e^{\theta_{kji}} \right) \right] \\
&+ \sum_{i=1}^{60} n_i \log [f(\boldsymbol{\rho}, \mathbf{u}_i)] \\
&- \sum_{i=1}^{60} n_i \log \left[ \sum_{i=1}^{30} \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \theta_{kji}}}{(1 + \sum_{j=1}^3 e^{\theta_{kji}})} f(\boldsymbol{\rho}, \mathbf{u}_i) \right]. \quad (3.30)
\end{aligned}$$

Finalmente, ao maximizar  $\ell(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$  sob as restrições R1 e R2, obtém-se os estimadores de máxima verossimilhança para os parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho})$ .

### 3.2.1 Modelo de Dependência de Locação

Os modelos baseados na representação de Bahadur para o caso multinomial consideram apenas as correlações dois a dois, diferente do modelo para o caso binomial que apresenta uma medida de correlação que relaciona as três posições do códon simultaneamente. Essa decisão foi baseada no número de parâmetros desses modelos, que como será visto nos modelos seguintes, já é muito grande apenas com as correlações dois a dois. Apesar de este primeiro modelo apresentado ter apenas três parâmetros de correlação dois a dois e portanto o número de parâmetros não ser um problema, para que todos os modelos sejam coerentes entre si, nenhum deles tem a correlação que relaciona as três posições.

O modelo de dependência de locação considera apenas a correlação entre a mudança de uma posição do códon para as seguintes, sem levar em conta qual a base nitrogenada das posições. Ou seja, a estrutura de dependência de se ter uma passagem de uma base A na posição 1 para uma base T na posição 2, por exemplo, é a mesma de uma passagem de uma base C na posição 1 para uma base G na posição 2. Assim, a estrutura de dependência  $f(\boldsymbol{\rho}, \mathbf{u}_i)$  é:

$$f(\boldsymbol{\rho}, \mathbf{u}_i) = 1 + \sum_{j=1}^3 \sum_{s=1}^3 (\rho_{12} u_{1ji} u_{2si} + \rho_{13} u_{1ji} u_{3si} + \rho_{23} u_{2ji} u_{3si}), \quad (3.31)$$



em que  $\rho_{12} = E(U_{1ji}U_{2si})$ ,  $\rho_{13} = E(U_{1ji}U_{3si})$  e  $\rho_{23} = E(U_{2ji}U_{3si})$ . Somando os 3 parâmetros  $\boldsymbol{\rho}$  de correlação aos 12 parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  da parte independente, esse modelo tem um total de 15 parâmetros.

### 3.2.2 Modelo de Dependência de Transição

Esse modelo considera apenas a correlação entre a mudança de uma base para outra, independente da posição dentro do códon, ou seja, a correlação de uma mudança de base A na posição 1 para a base T na base 2, por exemplo, é a mesma de uma mudança da base A na posição 1 para a base T na posição 3. Assim, a estrutura de dependência  $f(\boldsymbol{\rho}, \mathbf{u}_i)$  é:

$$f(\boldsymbol{\rho}, \mathbf{u}_i) = 1 + \sum_{j=1}^3 \sum_{s=1}^3 \rho_{js} (u_{1ji}u_{2si} + u_{1ji}u_{3si} + u_{2ji}u_{3si}), \quad (3.32)$$

em que  $\rho_{js} = E(U_{kji}U_{lsi})$ ,  $\forall k \neq l$ . Somando os 9 parâmetros  $\boldsymbol{\rho}$  de correlação aos 12 parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  da parte independente, esse modelo tem um total de 21 parâmetros.

### 3.2.3 Modelo de Dependência de Semi-Locação e Transição

Esse modelo considera a correlação entre a mudança de uma determinada base para outra, assim como o modelo de dependência de transição. No entanto, leva em conta a distância da transição dentro do códon, ou seja, assume que a correlação de uma mudança da posição 1 para 2 e de uma mudança da posição 2 para 3 é a mesma, mas essa correlação é diferente daquela de uma mudança da posição 1 para 3. Esse modelo é chamado de dependência de semi-locação, pois ainda considera as correlações de mudanças da posição 1 para 2 e 2 para 3 iguais. Assim, a estrutura de dependência  $f(\boldsymbol{\rho}, \mathbf{u}_i)$  é:

$$f(\boldsymbol{\rho}, \mathbf{u}_i) = 1 + \sum_{j=1}^3 \sum_{s=1}^3 [\rho_{1,js} (u_{1ji}u_{2si} + u_{2ji}u_{3si}) + \rho_{2,js} u_{1ji}u_{3si}], \quad (3.33)$$

em que  $\rho_{1,js} = E(U_{1ji}U_{2si}) = E(U_{2ji}U_{3si})$  e  $\rho_{2,js} = E(U_{1ji}U_{3si})$ . Somando os 18 parâmetros  $\boldsymbol{\rho}$  de correlação aos 12 parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  da parte independente, esse mo-

delo tem um total de 30 parâmetros.

### 3.2.4 Modelo de Dependência de Locação e Transição

Esse modelo é considerado o modelo de dependência completa, pois leva em conta a correlação de mudança de cada uma das posições do códon para as demais, e de cada mudança de base separadamente. Ou seja, a correlação de uma mudança da base A na posição 1 para a base T na posição 2, por exemplo, é diferente da mesma mudança de bases da posição 2 para 3; a correlação de uma mudança da base A para a base T também será diferente, para todas as posições do códon. Assim, a estrutura de dependência  $f(\boldsymbol{\rho}, \mathbf{u}_i)$  é:

$$f(\boldsymbol{\rho}, \mathbf{u}_i) = 1 + \sum_{j=1}^3 \sum_{s=1}^3 (\rho_{1j,2s} u_{1ji} u_{2si} + \rho_{1j,3s} u_{1ji} u_{3si} + \rho_{2j,3s} u_{2ji} u_{3si}), \quad (3.34)$$

em que  $\rho_{1j,2s} = E(U_{1ji}U_{2si})$ ,  $\rho_{1j,3s} = E(U_{1ji}U_{3si})$  e  $\rho_{2j,3s} = E(U_{2ji}U_{3si})$ . Somando os 27 parâmetros  $\boldsymbol{\rho}$  de correlação aos 12 parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  da parte independente, esse modelo tem um total de 39 parâmetros.

### 3.2.5 Método de Estimação dos Parâmetros

Nos modelos baseados na representação de Bahadur, para o caso de respostas politômicas, a maximização da log-verossimilhança sob as duas restrições para estimar os parâmetros simultaneamente também é um problema computacionalmente complexo, a estimação é feita em dois passos, assim como no modelo de respostas binárias. O primeiro passo estima  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  através do modelo independente descrito na seção de modelos regressivos.

Em seguida, utilizando  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$  estimados, obtém-se  $\hat{\theta}_{kji}$  e  $\hat{\pi}_{kji}$  para todo  $k, j = 1, 2, 3$  e  $i = 1, \dots, 60$ . A partir dessas probabilidades, calcula-se a probabilidade dos códons assumindo independência entre as posições,  $\hat{P}_I(\mathbf{Y}_i | \mathbf{X}_i)$ .

As variáveis para a estrutura de dependência também são calculadas a partir das estimativas  $\hat{\pi}_{kji}$ :

$$\hat{U}_{kji} = \frac{Z_{kji} - \hat{\pi}_{kji}}{\sqrt{\hat{\pi}_{kji}(1 - \hat{\pi}_{kji})}}. \quad (3.35)$$

Finalmente, no segundo passo estima-se  $\boldsymbol{\rho} | (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ . Como  $\theta_{kji}$  e  $\pi_{kji}$  já foram estimados, o primeiro termo da equação (3.30) é constante com relação a  $\boldsymbol{\rho}$ , portanto a log-verossimilhança a ser maximizada é

$$\ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left\{ \log [f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)] - \log \left[ \sum_{i=1}^{30} \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \hat{\theta}_{kji}}}{(1 + \sum_{j=1}^3 e^{\hat{\theta}_{kji}})} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right] \right\}, \quad (3.36)$$

assim, para obter as estimativas de máxima verossimilhança, basta resolver o sistema de equações a seguir:

$$\left\{ \frac{\partial}{\partial \rho_*} \ell(\boldsymbol{\rho}) = 0, \quad \forall \rho_* \in \boldsymbol{\rho}. \right. \quad (3.37)$$

### 3.2.6 Gradiente e Informação de Fisher

Cada elemento do vetor gradiente é dado por,

$$\frac{\partial}{\partial \rho_*} \ell(\boldsymbol{\rho}) = \sum_{i=1}^{60} n_i \left[ \frac{1}{f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i)} \frac{\partial}{\partial \rho_*} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) - \frac{\sum_{i=1}^{60} \frac{\partial}{\partial \rho_*} P_B(\mathbf{Y}_i | \mathbf{X}_i)}{\sum_{i=1}^{60} P_B(\mathbf{Y}_i | \mathbf{X}_i)} \right], \quad (3.38)$$

para todo  $\rho_* \in \boldsymbol{\rho}$ .

Relembrando que pela equação (3.28), e considerando que tem-se os valores estimados dos parâmetros  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ , tem-se que,

$$P_B(\mathbf{Y}_i | \mathbf{X}_i) = \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \hat{\theta}_{kji}}}{(1 + \sum_{j=1}^3 e^{\hat{\theta}_{kji}})} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i), \quad (3.39)$$

então,

$$\frac{\partial}{\partial \rho_*} P_B(\mathbf{Y}_i | \mathbf{X}_i) = \prod_{k=1}^3 \frac{e^{\sum_{j=1}^3 z_{kji} \hat{\theta}_{kji}}}{(1 + \sum_{j=1}^3 e^{\hat{\theta}_{kji}})} \frac{\partial}{\partial \rho_*} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i). \quad (3.40)$$

A matriz de informação de Fisher é dada por

$$I_F(\boldsymbol{\rho}) = -\mathbb{E} \left( \frac{\partial^2}{\partial \rho_{*1} \partial \rho_{*2}} \ell(\boldsymbol{\rho}) \right), \quad (3.41)$$

tal que, para todo  $\rho_{*1}, \rho_{*2} \in \boldsymbol{\rho}$ ,

$$\begin{aligned} \frac{\partial^2}{\partial \rho_{*1} \partial \rho_{*2}} \ell(\boldsymbol{\rho}) &= - \sum_{i=1}^{60} n_i \frac{\left[ \frac{\partial}{\partial \rho_{*1}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right] \left[ \frac{\partial}{\partial \rho_{*2}} f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right]}{\left[ f(\boldsymbol{\rho}, \hat{\mathbf{u}}_i) \right]^2} \\ &+ \sum_{i=1}^{60} n_i \frac{\left[ \sum_{i=1}^{60} \frac{\partial}{\partial \rho_{*1}} P_B(\mathbf{Y}_i | \mathbf{X}_i) \right] \left[ \sum_{i=1}^{60} \frac{\partial}{\partial \rho_{*2}} P_B(\mathbf{Y}_i | \mathbf{X}_i) \right]}{\left[ \sum_{i=1}^{60} P_B(\mathbf{Y}_i | \mathbf{X}_i) \right]^2}. \end{aligned} \quad (3.42)$$

Portanto, os elementos da matriz de informação de Fisher são,

$$\begin{aligned}
 -\mathbb{E}\left(\frac{\partial^2}{\partial\rho_{*1}\partial\rho_{*2}}\ell(\boldsymbol{\rho})\right) &= -\sum_{i=1}^{60}n_i\mathbb{E}\left(\frac{\left[\sum_{i=1}^{60}\frac{\partial}{\partial\rho_{*1}}P_B(\mathbf{Y}_i|\mathbf{X}_i)\right]\left[\sum_{i=1}^{60}\frac{\partial}{\partial\rho_{*2}}P_B(\mathbf{Y}_i|\mathbf{X}_i)\right]}{\left[\sum_{i=1}^{60}P_B(\mathbf{Y}_i|\mathbf{X}_i)\right]^2}\right) \\
 &+ \sum_{i=1}^{60}n_i\frac{\left[\frac{\partial}{\partial\rho_{*1}}f(\boldsymbol{\rho},\hat{\mathbf{u}}_i)\right]\left[\frac{\partial}{\partial\rho_{*2}}f(\boldsymbol{\rho},\hat{\mathbf{u}}_i)\right]}{\left[f(\boldsymbol{\rho},\hat{\mathbf{u}}_i)\right]^2}. \tag{3.43}
 \end{aligned}$$

# Capítulo 4

## Medidas de Ajuste dos Modelos

Para que os modelos propostos e descritos, tanto para respostas binárias quanto para respostas politômicas, sejam avaliados com relação ao ajuste, é necessário que se aplique algumas técnicas que permitam essa avaliação, bem como comparação dos resultados obtidos para cada modelo. Esse capítulo trata, portanto, de medidas de ajuste de modelos aplicáveis ao que é proposto neste trabalho, e também de métodos de validação dos modelos.

### 4.1 Soma de Quadrado dos Erros (SQE)

A SQE deve ser a menor possível, pois quanto mais baixo o valor obtido para essa medida, mais próximos os valores estimados estão dos verdadeiros, e é obtida através da expressão

$$SQE = \sum_{i=1}^{60} \left[ P(\mathbf{Y}_i | \mathbf{X}_i) - \hat{P}(\mathbf{Y}_i | \mathbf{X}_i) \right]^2. \quad (4.1)$$

### 4.2 Critério de Informação de Akaike (AIC)

O AIC ([Akaike, 1973](#)) leva em consideração a log-verossimilhança do modelo ajustado e o número de parâmetros desses modelos. Quanto menor o AIC, melhor o ajuste obtido,

pois esse critério prioriza modelos com maior verossimilhança, mas penaliza modelos que possuem parâmetros em excesso, ou seja, é uma medida que leva em consideração um balanceamento entre máxima verossimilhança e total de parâmetros estimados. Esse critério é dado pela seguinte expressão:

$$AIC = 2 \times [\text{número de parâmetros} - \log(\text{verossimilhança})]. \quad (4.2)$$

### 4.3 Critério de Informação Bayesiano (BIC)

O BIC (Raftery, 1986) também leva em consideração a log-verossimilhança do modelo ajustado e o número de parâmetros desse modelos. Assim como o AIC, quanto menor o BIC, melhor o ajuste obtido.

A diferença entre o AIC e o BIC, é que o segundo penaliza os modelos pelo número de parâmetros, ponderando essa penalização pelo logaritmo do tamanho da amostra. Esse critério é dado pela seguinte expressão:

$$BIC = \log(N) \times (\text{número de parâmetros}) - 2 \times \log(\text{verossimilhança}). \quad (4.3)$$

### 4.4 Função Desvio

A função desvio é muito utilizada na análise de modelos lineares generalizados, e leva em consideração a diferença entre a log-verossimilhança do modelo saturado (calculada sobre os valores observados) e a log-verossimilhança do modelo estimado, sendo então dada pela seguinte expressão,

$$D(\mathbf{p}; \hat{\boldsymbol{\pi}}) = -2 [\ell(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}) - \ell(\boldsymbol{\pi}, \boldsymbol{\pi})]. \quad (4.4)$$

### 4.5 Validação Cruzada

A validação cruzada é um método estatístico para avaliar ou comparar a performance de um ou mais modelos ou algoritmos aplicados em uma amostra, separando-a em dois

grupos, um para ajuste (grupo de treino) e outro para validação (grupo de validação). Há diversas maneiras de fazer a validação cruzada em um banco de dados; a mais usual é a técnica K-dobras e as demais formas de validação cruzada são variações dela.

### 4.5.1 K-dobras

Nessa técnica uma amostra é particionada em  $K$  grupos aleatórios de tamanhos iguais (ou quase iguais), e  $K$  iterações são realizadas da seguinte maneira: um grupo é separado para validação e o modelo é ajustado para os  $K - 1$  demais grupos. Em seguida o ajuste obtido é comparado com os valores do grupo de validação.

### 4.5.2 *Hold-out*

A validação *hold-out* separa o conjunto de dados em dois grupos aleatórios de tamanhos iguais, um grupo de treino e o outro de validação. Em seguida, o modelo é ajustado para o grupo de treino e o resultado comparado com o grupo de validação. Esse tipo de validação produz resultados muito dependentes da escolha dos grupos de teste e validação, por isso esse grupos devem ser aleatorizados, para que não haja tanto efeito da separação dos dados. Outra questão importante, é que para amostras pequenas a validação *hold-out* não é indicada, pois não apenas o efeito da separação dos dados será maior, como também haverá poucas observações nos grupos de treino e de validação.

### 4.5.3 *Leave-one-out*

A técnica *leave-one-out* é uma variação direta da K-dobras, em que o número  $K$  de grupos é igual ao tamanho da amostra, ou seja, para cada valor observado é feita uma validação. A vantagem dessa técnica é que a precisão das estimativas é praticamente não-viesada. A variância, entretanto, será maior do que a obtida se houver menos grupos. Outro problema dessa técnica é o custo, quanto maior a amostra, mais cara é a validação cruzada *leave-one-out*.

#### 4.5.4 K-dobras repetido

Essa validação realiza a técnica K-dobras repetidas vezes. Determina-se o número de repetições desejadas e para cada repetição, a técnica K-dobras é realizada completamente. Ao iniciar uma nova repetição, os grupos são realeatorizados.

### 4.6 Teste da Razão de Verossimilhança

O teste da razão de verossimilhança é feito para avaliar se a hipótese nula de que um parâmetro ou um conjunto de parâmetros sejam iguais a zero é verdadeira. Seja um modelo  $M$  cujos parâmetros são representados por  $\boldsymbol{\beta}$ ;  $M_0$  é este modelo sob  $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$  e  $L_0$  a sua verossimilhança, e  $M_1$  é este modelo sob  $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$  e  $L_1$  a sua verossimilhança. Assintoticamente, têm-se que

$$G^2 = -2 \log \left( \frac{L_0}{L_1} \right) \sim \chi_{p_1 - p_0}^2, \quad (4.5)$$

em que  $p_0 \leq p_1$  denotam o número de parâmetros dos modelos  $M_0$  e  $M_1$ , respectivamente.

### 4.7 Teste de Wald

O teste de Wald pode ser usado para testar  $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ , para um conjunto de parâmetros  $\boldsymbol{\beta}$  de um determinado modelo. Assim, a estatística do teste é dada por,

$$W_C = \hat{\boldsymbol{\beta}}' \mathbf{C}' \left\{ \mathbf{C} [\mathbf{Var}(\hat{\boldsymbol{\beta}})]^{-1} \mathbf{C}' \right\}^{-1} \mathbf{C} \hat{\boldsymbol{\beta}} \sim \chi_c^2, \quad (4.6)$$

em que  $c$  é o total de restrições determinadas por  $H_0$ .

Quando  $\hat{\boldsymbol{\beta}}$  são estimadores de máxima verossimilhança de  $\boldsymbol{\beta}$ , sabe-se que  $\mathbf{Var}(\hat{\boldsymbol{\beta}}) \xrightarrow{P} I_F^{-1}(\boldsymbol{\beta})$ , e nesse caso, a estatística do teste de Wald é dada por,

$$W_C = \hat{\boldsymbol{\beta}}' \mathbf{C}' [\mathbf{C} I_F(\boldsymbol{\beta}) \mathbf{C}']^{-1} \mathbf{C} \hat{\boldsymbol{\beta}} \sim \chi_c^2, \quad (4.7)$$

em que  $c$  é o total de restrições determinadas por  $H_0$ .



## 4.8 Teste de Escore

Quando a matriz de informação de Fisher não pode ser calculada, muitas vezes devido à esperança de  $\frac{\partial^2}{\partial\beta_i\partial\beta_j}\ell(\boldsymbol{\beta})$  não ter valor fechado, o teste de escore é adequado, pois utiliza a informação de Fisher observada, e a estatística do teste para  $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$  é dada por,

$$S = \mathbf{U}' \hat{I}_F^{-1}(\boldsymbol{\beta}) \mathbf{U} \sim \chi_c^2, \quad (4.8)$$

em que  $c$  é o total de restrições determinadas por  $H_0$  e,

$$\mathbf{U} = \frac{\partial}{\partial\beta_i}\ell(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}, \quad (4.9)$$

$$\hat{I}_F(\boldsymbol{\beta}) = \frac{\partial^2}{\partial\beta_i\partial\beta_j}\ell(\boldsymbol{\beta}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}. \quad (4.10)$$

# Capítulo 5

## Aplicação e Resultados

### 5.1 Implementação Computacional

Todas as implementações dos modelos e métodos propostos foram realizadas utilizando o software R ([R-project, 2010](#)), através de rotinas já contidas em pacotes do software, e outras programadas para executar exatamente os modelos propostos e suas análises necessárias.

As log-verossimilhanças foram programadas como funções, e para se obter os estimadores de máxima verossimilhança dos parâmetros nos modelos regressivos utilizou-se a rotina de otimização *optim*, com o método numérico de Broyden-Fletcher-Goldfarb-Shanno (BFGS). Nos modelos baseados na representação de Bahadur, utilizou-se a rotina de otimização *constrOptim*, também com o método numérico de BFGS. A diferença entre essas duas rotinas é que a primeira estima livremente os parâmetros, e a segunda permite inserir restrições não-lineares sobre os parâmetros, permitindo que (R1) e (R2) dos modelos baseados na representação de Bahadur sejam respeitadas.

A implementação computacional dos modelos foi uma das etapas mais importantes desse trabalho, devido aos inúmeros problemas e soluções que surgiram envolvendo os modelos baseados na representação de Bahadur. O primeiro problema encontrado foi estimar simultaneamente todos os parâmetros  $(\alpha, \beta, \rho)$  dos modelos baseados na repre-

sentação de Bahadur. A restrição  $f(\boldsymbol{\rho}, \mathbf{U}_i) > 0$  tornou difícil a estimação simultânea dos parâmetros  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$  relacionados ao logito (e às covariáveis) e dos parâmetros  $\boldsymbol{\rho}$  de correlação. Por isso foi tomada a decisão de estimá-los em duas etapas, primeiramente assumindo independência, obtendo  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}|\boldsymbol{\rho} = \mathbf{0})$ , e em seguida estimando  $(\hat{\boldsymbol{\rho}}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ .

Em uma primeira tentativa, tentou-se uma solução para estimar  $(\hat{\boldsymbol{\rho}}|\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$  através de uma regressão linear. Após estimados os parâmetros  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}|\boldsymbol{\rho} = \mathbf{0})$ , as probabilidades sob o modelo independente eram calculadas, bem como as variáveis  $\mathbf{u}_i$ , e a função  $f(\boldsymbol{\rho}, \mathbf{U}_i)$  era calculada como um fator de correção necessário para que as probabilidades estimadas pelo modelo baseado na representação de Bahadur apresentassem erro mínimo quando comparadas com as probabilidades observadas. Assim, para se obter as correlações, bastava realizar uma regressão linear, com  $f(\boldsymbol{\rho}, \mathbf{U}_i)$  (como fator de correção) sendo a variável resposta, os produtos das variáveis  $\mathbf{u}_i$  (de acordo com cada um dos modelos propostos) sendo as preditoras, e as correlações  $\boldsymbol{\rho}$  sendo os parâmetros a serem estimados. No entanto, para a realização dessa regressão linear, era necessário suposições de normalidade, o que verificou-se não ser verdadeira, e esse método foi portanto descartado.

A estimação em dois passos no entanto continua sendo até o momento a melhor solução para os problemas computacionais, com o segundo passo da estimação através de máxima verossimilhança. Pesquisas sobre o software utilizado levaram à rotina *constrOptim*, que permite a estimação desse segundo passo, incluindo as restrições necessárias para que os modelos estimem as probabilidades corretamente, respeitando todas as leis probabilísticas.

## 5.2 Modelos Binomiais

A Tabela 5.1 apresenta os resultados obtidos para os modelos binomiais multivariados, valores de AIC, BIC, função desvio e SQE. A partir das medidas de ajuste obtidas, é possível constatar que dentre os modelos com respostas binárias, para qualquer uma das funções de ligação utilizada, o melhor modelo é o aditivo, e dentre as funções de

ligação, a que tem melhor ajuste é o logito.

Tabela 5.1: Medidas dos Modelos Binomiais

Ligação	Modelo	AIC	BIC	Função Desvio	SQE	Par.
Logito	Independente	110316,6	110361,7	12126,75	0,0168	6
	Ig. Preditivo	110245,7	110298,4	12053,90	0,0166	7
	Markov	110213,1	110273,4	12019,30	0,0165	8
	Aditivo	<b>110183,6</b>	<b>110251,4</b>	<b>11987,81</b>	<b>0,0165</b>	9
	Bahadur	110264,5	110339,8	12066,69	0,0167	10
Probit	Independente	110374,7	110419,9	12184,92	0,0168	6
	Ig. Preditivo	110375,2	110427,9	12183,41	0,0168	7
	Markov	<b>110317,6</b>	<b>110377,8</b>	12123,77	0,0167	8
	Aditivo	110318,6	110386,3	<b>12122,74</b>	<b>0,0167</b>	9
	Bahadur	110353,8	110429,1	12156,01	0,0168	10
Log-Log	Independente	110394,5	110439,7	12204,66	0,0168	6
Compl.	Ig. Preditivo	110331,2	110383,9	12139,38	0,0167	7
	Markov	110301,3	110361,6	12107,53	0,0167	8
	Aditivo	<b>110273,5</b>	<b>110341,3</b>	<b>12077,72</b>	<b>0,0166</b>	9
	Bahadur	110353,3	110428,6	12155,50	0,0168	10

As Figuras 5.1 e 5.2 mostram os ajustes das probabilidades para os modelos aditivo e de Bahadur, com função de ligação logito. Pode-se observar, a partir do gráfico, que o ajuste dos modelos binomiais não é tão bom quanto desejado. Note que há alguns códons com valores observados bem maiores que os estimados

### 5.3 Modelos Multinomiais

As Tabelas 5.2 e 5.3 apresentam os resultados obtidos para os modelos multinomiais multivariados, valores de AIC, BIC, função desvio e SQE. A partir das medidas de ajuste obtidas, é possível constatar que dentre os modelos regressivos com respostas

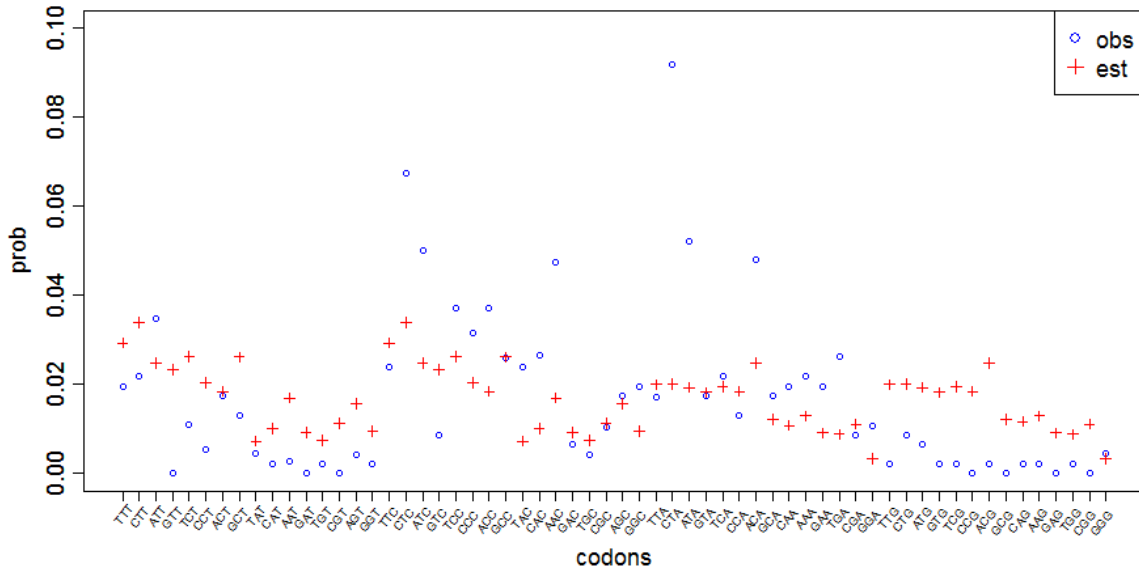


Figura 5.1: Valores observados *versus* estimados para os códons do modelo logístico regressivo aditivo com ligação logito

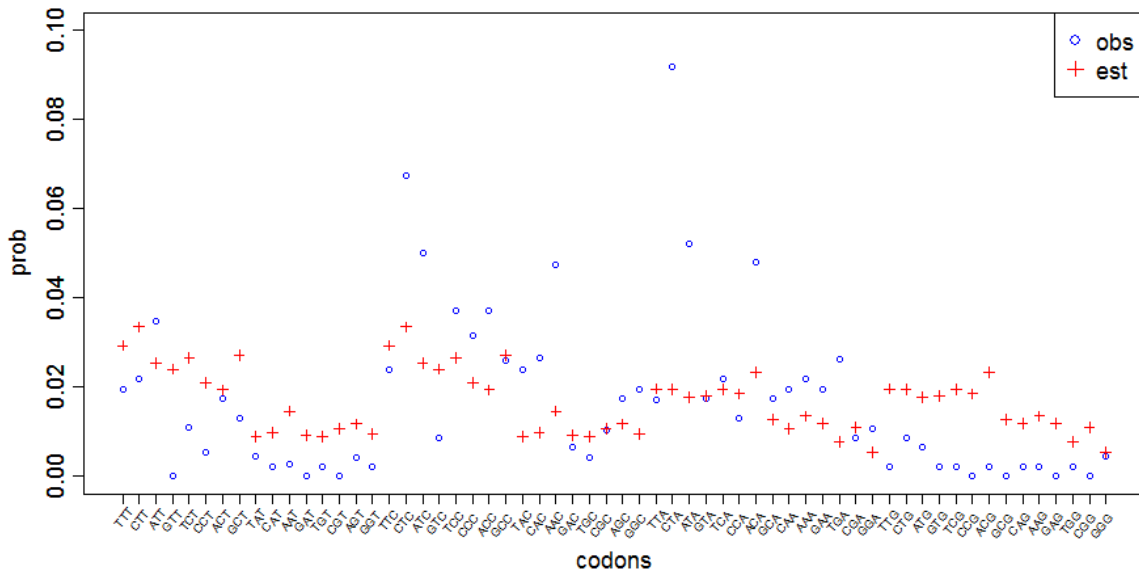


Figura 5.2: Valores observados *versus* estimados para os códons do modelo baseado na representação de Bahadur com ligação logito

politômicas, o melhor modelo é o aditivo, e dentre os modelos baseados na representação de Bahadur, o melhor é o modelo de semi-locação & transição.

Tabela 5.2: Medidas dos Modelos Multinomiais Regressivos

Modelo	AIC	BIC	Função Desvio	SQE	Par.
Independente	101723,21	101813,57	3521,39	0,0052	12
Ig. Preditivo	100939,08	101052,03	2731,26	0,0037	15
Markov	99766,62	99902,16	1552,80	0,0019	18
Aditivo	<b>99596,07</b>	<b>99754,21</b>	<b>1376,26</b>	<b>0,0017</b>	21

Tabela 5.3: Medidas dos Modelos Multinomiais Baseados na Representação de Bahadur

Modelo	AIC	BIC	Função Desvio	SQE	Par.
Locação	101588,80	101701,75	3380,98	0,0048	15
Transição	100740,39	100898,52	2520,57	0,0034	21
Semi-Loc. e Trans.	<b>100030,35</b>	<b>100256,26</b>	<b>1792,53</b>	<b>0,0024</b>	30
Locação e Transição	100144,70	100438,38	1888,89	0,0027	39

As Figuras 5.3 e 5.4 mostram os ajustes das probabilidades para os modelos aditivo e semi-locação e transição. É visível, a partir dos gráficos dos modelos multinomiais aditivo e semi-locação e transição, que o ajuste deles é superior ao dos modelos binomiais. Isso se deve ao fato de que classificar as bases como *purinas* ou *pirimidinas* resulta em uma perda da informação.

É interessante notar que não há apenas uma diferença gráfica entre os modelos binomiais e multinomiais, mas também nos valores obtidos para as medidas de ajuste desses modelos, principalmente na função desvio, que é muito menor nos modelos multinomiais, ou seja, a log-verossimilhança desses modelos é muito mais próxima da log-verossimilhança do modelo saturado, do que a dos modelos binomiais.

Comparando os gráficos dos modelos aditivo e de semi-locação e transição, não é possível afirmar que há grandes diferenças entre os dois ajustes. Há inclusive códons

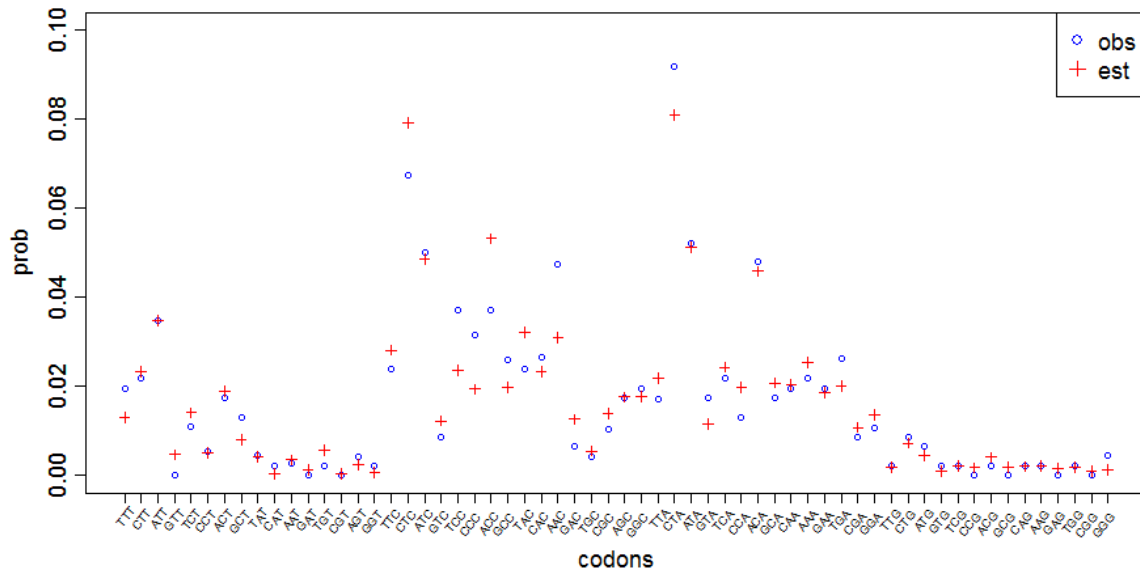


Figura 5.3: Valores observados *versus* estimados do modelo logístico regressivo aditivo

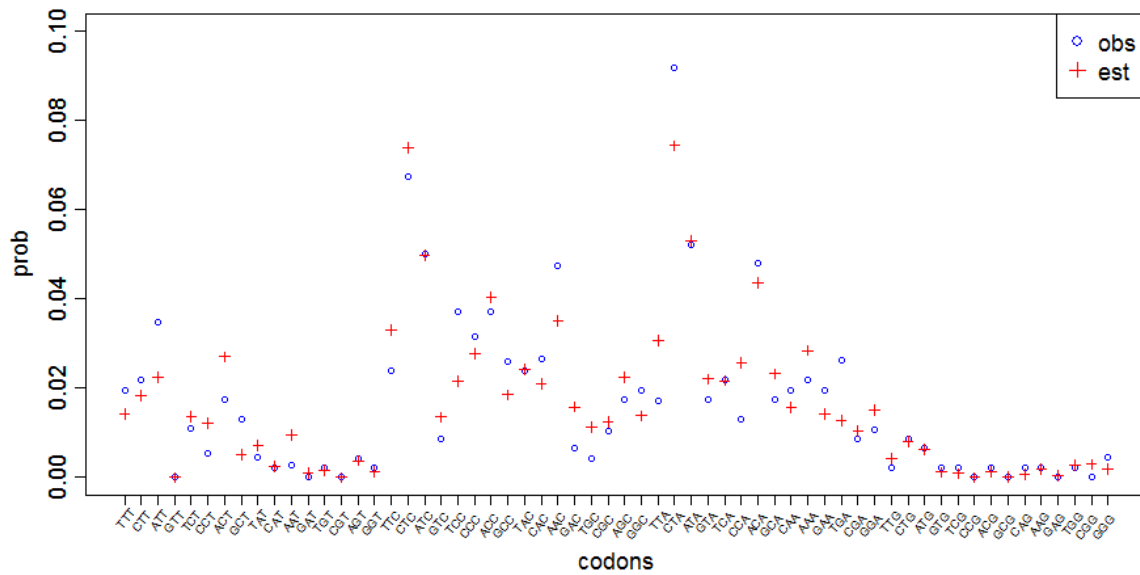


Figura 5.4: Valores observados *versus* estimados do modelo baseado na representação de Bahadur com dependência de semi-locção e transição

Tabela 5.4: Estimativas dos Parâmetros do Modelo Logístico Regressivo Aditivo

$\alpha_{11}$	=	3,6827	$\gamma_{11}$	=	-1,6219	$\beta_1$	=	-0,0231
$\alpha_{12}$	=	4,0360	$\gamma_{12}$	=	-0,2981	$\beta_2$	=	-0,2183
$\alpha_{13}$	=	3,1445	$\gamma_{13}$	=	0,3719	$\beta_3$	=	-0,4768
$\alpha_{21}$	=	3,1393	$\gamma_{21}$	=	-0,0732			
$\alpha_{22}$	=	3,1667	$\gamma_{22}$	=	-0,7823			
$\alpha_{23}$	=	3,1722	$\gamma_{23}$	=	-0,2868			
$\alpha_{31}$	=	4,2023	$\gamma_{31}$	=	0,1442			
$\alpha_{32}$	=	4,2246	$\gamma_{32}$	=	2,6228			
$\alpha_{33}$	=	1,8168	$\gamma_{33}$	=	3,0027			

melhor ajustados em cada um dos modelos, como por exemplo os códons CTT e ATT, que têm melhor ajuste no modelo aditivo, e os códons GTT e ACC, que têm melhor ajuste no modelo de semi-locação e transição. Nota-se também que há um melhor ajuste para as probabilidades de valores bem pequenos em ambos modelos.

As estimativas dos parâmetros para os modelos multinomiais aditivo e de dependência de semi-locação & transição encontram-se nas Tabelas 5.4 e 5.5, respectivamente.

No modelo aditivo, os parâmetros de dependência com maiores valores (em módulo) são  $\gamma_{11}$ ,  $\gamma_{32}$  e  $\gamma_{33}$ . Isso significa que a primeira posição do códon quando assume a base C, tem maior influência na *log-odds* da segunda posição do que quando assume as demais bases; também, a segunda posição do códon quando assume as bases A ou G, tem maior influência na *log-odds* da terceira posição do que quando assume a base C, e do que as bases da primeira posição.

No modelo de semi-locação e transição, as correlações com maiores valores (em módulo) são  $\rho_{1,11}$ ,  $\rho_{1,12}$ ,  $\rho_{1,13}$ ,  $\rho_{2,11}$ ,  $\rho_{2,12}$ ,  $\rho_{2,31}$  e  $\rho_{2,32}$ . Isso significa que há uma maior correlação entre as mudanças de uma base C para as demais, da posição 1 para 2 ou da posição 2 para 3, e para as mudanças de uma base C para C ou A, ou de uma base G para C ou A, da posição 1 para 3.

A validação cruzada foi aplicada para os modelos multinomiais, considerando como



Tabela 5.5: Estimativas dos Parâmetros do Modelo Baseado na Representação de Bahadur de Dependência de Semi-Localção e Transição

$\alpha_{11} = -1,0156$	$\rho_{1,11} = -0,1001$	$\rho_{2,11} = 0,1313$	$\beta_1 = -0,0062$
$\alpha_{12} = -0,8952$	$\rho_{1,12} = -0,1178$	$\rho_{2,12} = 0,1125$	$\beta_2 = 0,1801$
$\alpha_{13} = -1,8673$	$\rho_{1,13} = -0,1280$	$\rho_{2,13} = 0,0256$	$\beta_3 = 0,0836$
$\alpha_{21} = -1,7101$	$\rho_{1,21} = 0,0953$	$\rho_{2,21} = 0,0213$	
$\alpha_{22} = -2,1042$	$\rho_{1,22} = 0,0122$	$\rho_{2,22} = 0,0142$	
$\alpha_{23} = -2,4727$	$\rho_{1,23} = -0,0341$	$\rho_{2,23} = -0,0133$	
$\alpha_{31} = -0,1736$	$\rho_{1,31} = 0,0960$	$\rho_{2,31} = 0,1480$	
$\alpha_{32} = -0,2122$	$\rho_{1,32} = 0,0520$	$\rho_{2,32} = 0,1895$	
$\alpha_{33} = -2,6073$	$\rho_{1,33} = 0,0897$	$\rho_{2,33} = 0,0191$	

unidade amostral cada uma das 30 seqüências do gene NADH4. Devido a quantidade de seqüências, optou-se por usar a validação cruzada *leave-one-out*, assim, 30 iterações foram realizadas.

Para avaliar os resultados da validação cruzada, foi calculado o quadrado médio dos erros (QME) e a variância da SQE (Tabelas 5.6 e 5.7). Assim, como foi verificado pelas demais medidas de análise de ajuste (AIC, BIC e SQE), a validação cruzada também indica que, quando observa-se o QME das validações cruzadas, dentre os modelo regressivos o melhor deles é o modelo aditivo, e dentre os modelos baseados na representação de Bahadur o melhor deles é o modelo de semi-localção e transição. É interessante notar que as variâncias da SQE desses dois modelos não são as menores obtidas, quando comparadas com os demais modelos.

As Figuras 5.5 e 5.6 apresentam os gráficos da SQE das validações cruzadas para os modelos logísticos regressivos e para os modelos baseados na representação de Bahadur, respectivamente, em que as linhas tracejadas representam o QME da validação cruzada de cada modelo. É possível verificar visualmente que o modelo aditivo e o modelo de semi-localção e transição têm as menores médias dentre os modelos logísticos regressivos e os modelos baseados na representação de Bahadur, respectivamente. É visível também

Tabela 5.6: Resultados da Validação Cruzada dos Modelos Logísticos Regressivos

Modelo	$QME$	$Var(SQE)$
Independente	0,0052	<b><math>3,88 \times 10^{-13}</math></b>
Igualmente Preditivo	0,0036	$7,20 \times 10^{-13}$
Markov	0,0019	$1,04 \times 10^{-12}$
Aditivo	<b>0,0017</b>	$1,45 \times 10^{-12}$

Tabela 5.7: Resultados da Validação Cruzada dos Modelos Baseados na Representação de Bahadur

Modelo	$QME$	$Var(SQE)$
Locação	0,0048	<b><math>2,80 \times 10^{-13}</math></b>
Transição	0,0034	$1,32 \times 10^{-12}$
Semi-Locação e Transição	<b>0,0024</b>	$1,35 \times 10^{-9}$
Locação e Transição	0,0027	$7,95 \times 10^{-9}$

a maior variância da SQE no modelo de semi-locação e transição e no de locação e transição; apesar disso, os pontos aparentam estar homoganeamente dispersos entre si.

A Tabela 5.8 apresenta os resultados para os testes dos parâmetros de dependência e das covariáveis para o modelo aditivo de semi-locação e transição. Foi escolhido o teste da razão de verossimilhança para a realização dos testes de interesse. Em todos os testes, constatou-se que as covariáveis são significantes nos dois modelos, assim como os parâmetros de dependência também o são.

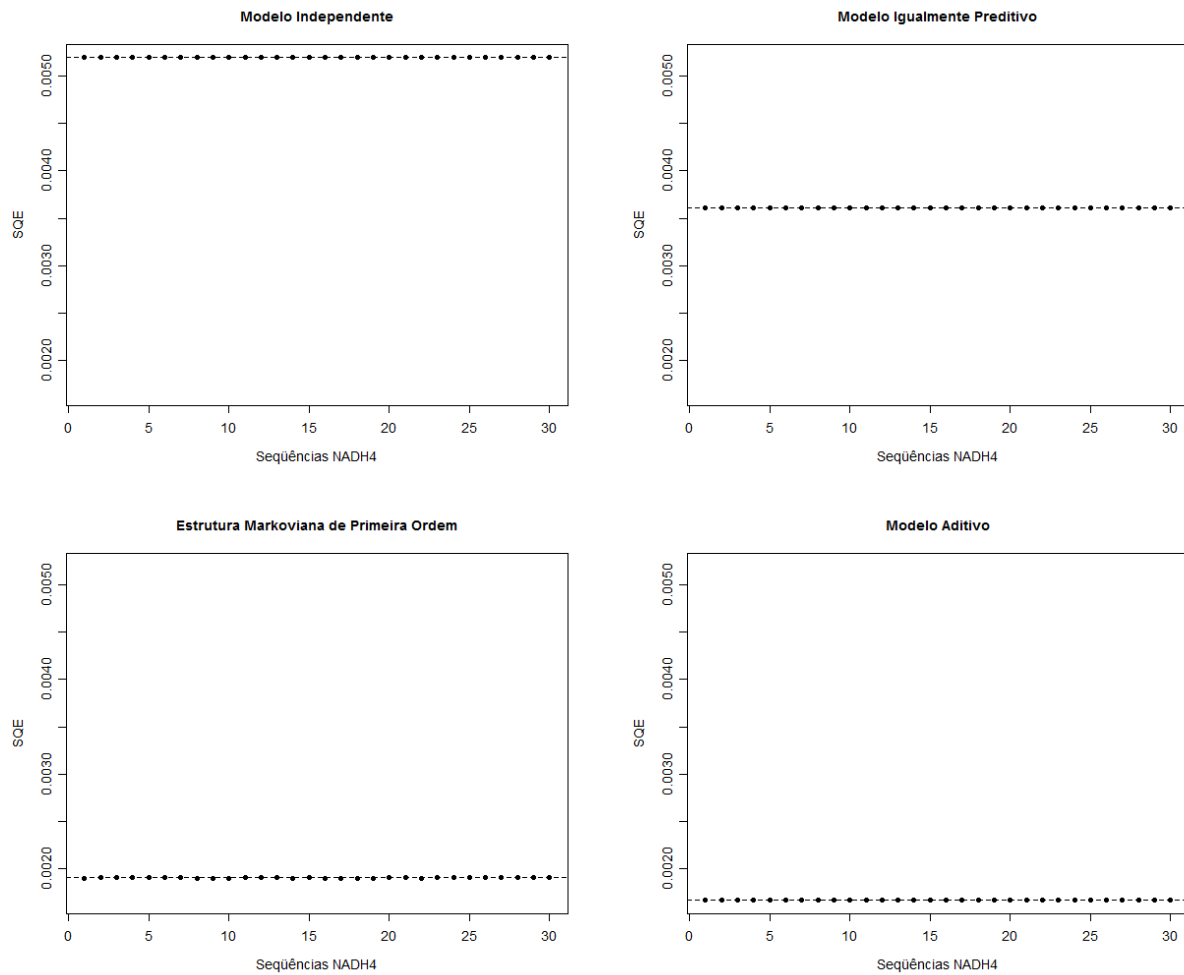


Figura 5.5: Validação Cruzada dos Modelos Logísticos Regressivos

Tabela 5.8: Testes dos Parâmetros do Modelo Aditivo e do Modelo de Semi-Localção &amp; Transição

Modelo	Teste	$G^2$	g.l.	p-valor
Aditivo	$H_0 : \gamma = \mathbf{0}$	2145,1385	9	< 0,0001
	$H_0 : \beta = \mathbf{0}$	239,9281	3	< 0,0001
Semi-Localção	$H_0 : \rho = \mathbf{0}$	1728,8617	18	< 0,0001
& Transição	$H_0 : \beta = \mathbf{0}$	110,7713	3	< 0,0001

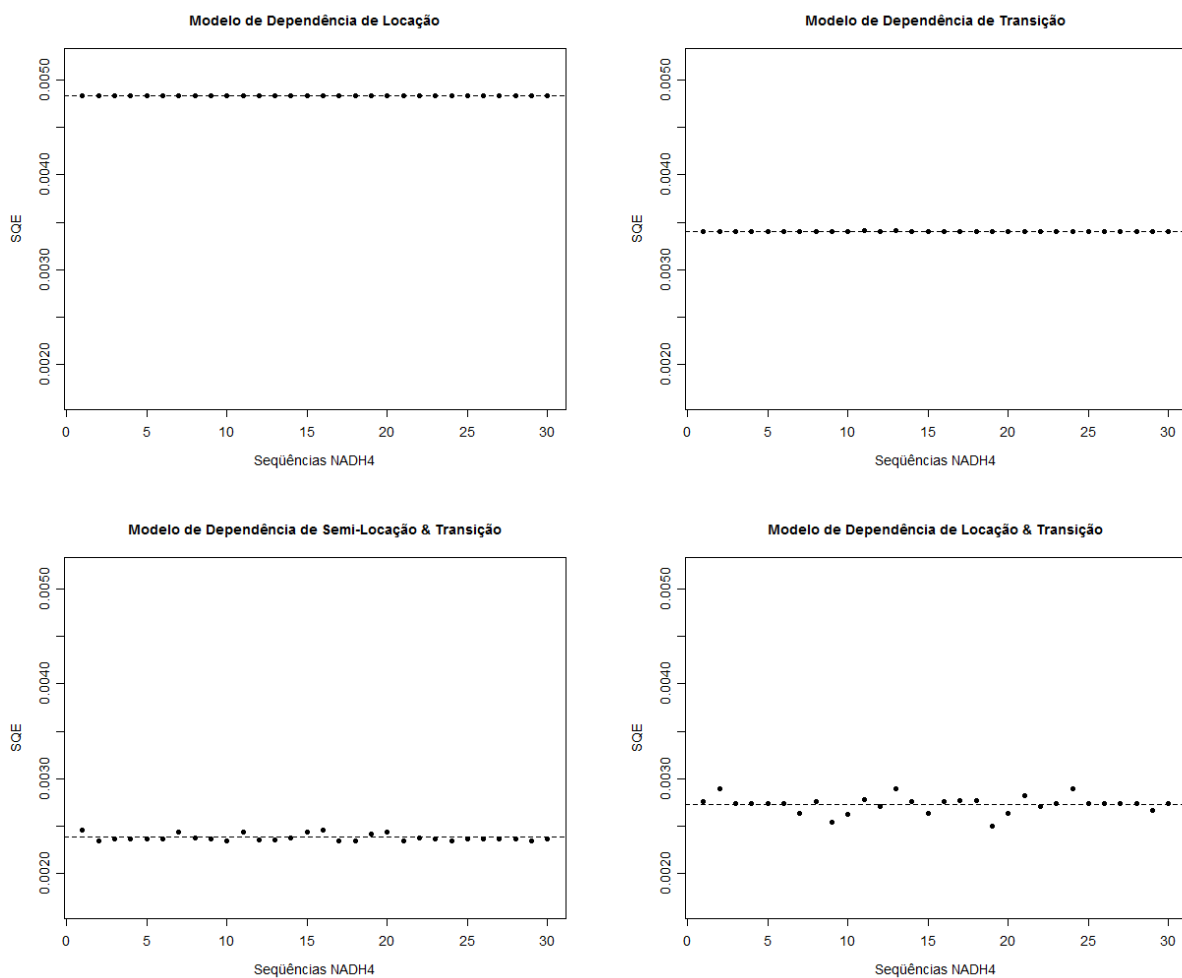


Figura 5.6: Validação Cruzada dos Modelos Baseados na Representação de Bahadur

## Capítulo 6

# Análise de Diagnóstico dos Modelos

Uma breve análise de diagnóstico dos modelos é feita nesse capítulo, com o intuito de fundamentar e validar os modelos propostos, indo além da análise de medidas e métodos de seleção de modelos, mas também verificando pontos de influência. Essas análises são feitas somente sobre os modelos multinomiais, por serem os modelos de melhor ajuste dos dados, quando comparados com os binomiais. Além disso, os diagnósticos foram realizados apenas para os modelos logísticos regressivos.

Os métodos de análise de pontos discrepantes, de influência e de alavanca foram primeiramente generalizados para modelos de regressão logística para dados binários (Pregibon, 1981) e estendidos para respostas múltiplas (Lesaffre e Albert, 1989) como um modelo linear generalizado multivariado. Seber e Nyangoma (2000) e Nyangoma et al. (2006) também abordam a análise de resíduos e diagnósticos de pontos influentes para dados com respostas multinomiais.

Os modelos baseados na representação de Bahadur, devido à complexidade da estrutura do modelo e da estimação em dois passos não têm essa análise, sendo isso uma proposta para trabalhos futuros.

## 6.1 Modelos Logísticos Regressivos

Nesses modelos, foram ajustados logitos que determinam a estrutura de dependência entre as três posições dos códons, considerando também três covariáveis. Os conjuntos de equações (3.10), (3.12), (3.14) e (3.16) dos modelos logísticos regressivos podem ser escritos matricialmente como,

$$\boldsymbol{\theta} = \mathbf{C} \log(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\xi}, \quad (6.1)$$

tal que,

$$\mathbf{C} = \left[ \begin{array}{c} \left( \begin{array}{cccc} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{array} \right) \otimes \mathbf{I}_{60} \\ \otimes \mathbf{I}_3 \end{array} \right], \quad (6.2)$$

o vetor de probabilidades,

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3)', \quad (6.3)$$

em que  $\boldsymbol{\pi}_k = (\boldsymbol{\pi}_{k,1}, \dots, \boldsymbol{\pi}_{k,60})'$  e  $\boldsymbol{\pi}_{k,i} = (\pi_{k0i}, \pi_{k1i}, \pi_{k2i}, \pi_{k3i})'$  para todo  $k = 1, 2, 3$  e  $i = 1, \dots, 60$ , e o vetor dos logitos,

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)', \quad (6.4)$$

em que  $\boldsymbol{\theta}_k = (\boldsymbol{\theta}_{k,1}, \dots, \boldsymbol{\theta}_{k,60})'$  e  $\boldsymbol{\theta}_{k,i} = (\theta_{k1i}, \theta_{k2i}, \theta_{k3i})'$  para todo  $k = 1, 2, 3$  e  $i = 1, \dots, 60$ .

Cada um dos quatro modelos descritos possui sua própria matriz  $\mathbf{X}$  de especificação e seu próprio conjunto de parâmetros  $\boldsymbol{\xi}$ , e por isso os diagnósticos são feitos separados para cada um deles, mas simultaneamente para as três posições dos códons.

Uma primeira medida de diagnóstico dos modelos, para verificar possíveis *outliers* é dada por

$$\chi_{k,i}^2 = \sum_{j=0}^3 \frac{n_i (z_{kji} - \hat{\pi}_{kji})^2}{\hat{\pi}_{kji}}, \quad \forall k = 1, 2, 3 \text{ e } i = 1, \dots, 60. \quad (6.5)$$

Valores altos de  $\chi_{k,i}^2$  sugerem um ajuste fraco. Assim, a estatística de “qualidade de ajuste” para cada posição do códon,  $\chi_k^2$  é definida,

$$\chi_k^2 = \sum_{i=1}^{60} \chi_{k,i}^2, \quad \forall k = 1, 2, 3. \quad (6.6)$$

A função desvio dada, quando se deseja analisar as probabilidades de cada posição do códon, pela estatística,

$$D_k = -2[\ell(\hat{\boldsymbol{\pi}}_k, \boldsymbol{\pi}_k) - \ell(\boldsymbol{\pi}_k, \boldsymbol{\pi}_k)], \quad (6.7)$$

e pode ser reescrita como  $D_k = \sum_{i=1}^{60} d_{k,i}^2$ , tal que cada  $d_{k,i}^2$  representa essa medida de concordância entre as log-verossimilhanças observadas e estimadas para cada ítem da amostra, no caso, cada códon.

O modelo linear generalizado para estimar os logits de respostas múltiplas das três posições do códon tem a matriz de projeção dada por,

$$\mathbf{H} = \hat{\boldsymbol{\Sigma}}^{1/2} \mathbf{X}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X})^{-1} \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{1/2}, \quad (6.8)$$

tal que,

$$\hat{\boldsymbol{\Sigma}} = \mathbf{diag}(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2, \hat{\boldsymbol{\Sigma}}_3) = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{\boldsymbol{\Sigma}}_3 \end{pmatrix}, \quad (6.9)$$

$$\hat{\boldsymbol{\Sigma}}_k = \mathbf{diag}(\hat{\boldsymbol{\Sigma}}_{k,1}, \dots, \hat{\boldsymbol{\Sigma}}_{k,60}) = \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{k,1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\boldsymbol{\Sigma}}_{k,60} \end{pmatrix}, \quad (6.10)$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{k,i} &= \mathbf{diag}(\boldsymbol{\pi}_{k,i}) - \boldsymbol{\pi}_{k,i}\boldsymbol{\pi}'_{k,i} \\ &= \begin{pmatrix} \pi_{k1i}(1 - \pi_{k1i}) & -\pi_{k1i}\pi_{k2i} & -\pi_{k1i}\pi_{k3i} \\ -\pi_{k1i}\pi_{k2i} & \pi_{k2i}(1 - \pi_{k2i}) & -\pi_{k2i}\pi_{k3i} \\ -\pi_{k1i}\pi_{k3i} & -\pi_{k2i}\pi_{k3i} & \pi_{k3i}(1 - \pi_{k3i}) \end{pmatrix}. \end{aligned} \quad (6.11)$$

É importante ressaltar que a matriz de variância  $\hat{\boldsymbol{\Sigma}}$  possui zeros onde estariam as covariâncias entre as posições do códon, pois a estrutura de dependência entre elas está inserida na matriz de especificação  $\mathbf{X}$ , como covariáveis do modelo. Também na matriz  $\hat{\boldsymbol{\Sigma}}_k$ , as covariâncias entre os códon é zero, pois neste trabalho os códon são considerados independentes entre si.

A variabilidade dos parâmetros estimados  $\hat{\boldsymbol{\xi}}$  é dada pelo volume do elipsóide de confiança assintótico para  $\boldsymbol{\xi}$ , dado por  $|(\mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X})^{-1}|^{1/2}$ . Quando a  $i$ -ésima observação é retirada, o volume do elipsóide é dado por  $|(\mathbf{X}'_{(i)}\hat{\boldsymbol{\Sigma}}_{(i)}\mathbf{X}_{(i)})^{-1}|^{1/2}$ , em que o índice  $(i)$  indica as matrizes de desenho e variância correspondentes, sem a  $i$ -ésima observação.

Seja também a matriz  $\mathbf{M} = \mathbf{I} - \mathbf{H}$ , uma matriz de blocos assim como a matriz  $\mathbf{H}$ , tal que cada bloco  $\mathbf{M}_{ij}$  ou  $\mathbf{H}_{ij}$  seja uma matriz  $3 \times 3$ , para todo  $i, j = 1, \dots, 60$ . É possível avaliar a influência da  $i$ -ésima observação através da medida,

$$\frac{|\mathbf{X}'_{(i)}\hat{\boldsymbol{\Sigma}}_{(i)}\mathbf{X}_{(i)}|}{|\mathbf{X}'\hat{\boldsymbol{\Sigma}}\mathbf{X}|} \approx |\mathbf{M}_{ii}|, \quad (6.12)$$

satisfazendo  $0 \leq |\mathbf{M}_{ii}| < 1$ , de forma que quando  $|\mathbf{M}_{ii}|$  tem valor próximo de zero, indica um possível impacto da  $i$ -ésima observação nos estimadores de máxima verossimilhança.

Como forma de estabilizar a variância de  $\chi_{k,i}$ , o “diagnóstico studentizado” de *outliers* para modelos lineares generalizados multinomiais é apresentado da seguinte forma,

$$\chi_{k,i}^* = [\mathbf{M}_{ii}]_{JJ}^{-1/2} \chi_{k,i}, \quad (6.13)$$

que possui matriz de variância-covariância aproximadamente igual à identidade, tal que  $J = \sum_{j=1}^3 j z_{kji}$  e  $[\mathbf{M}_{ii}]_{JJ}$  o  $J$ -ésimo ( $J=1,2,3$ ) elemento da diagonal da matriz  $3 \times 3$ ,  $\mathbf{M}_{ii}$ . É possível demonstrar que a distribuição assintótica de  $\chi_{k,i}^{*2}$  não é  $\chi^2$  (Lesafre e Albert, 1989).

Como medida geral de discrepância entre  $\hat{\boldsymbol{\xi}}$ , as estimativas dos parâmetros com todas as observações, e  $\hat{\boldsymbol{\xi}}_{(i)}$ , as estimativas dos parâmetros sem a  $i$ -ésima observação, pode-se utilizar a distância generalizada de Cook,

$$d_{kc(i)} = \chi_{k,i}^2 [\mathbf{M}_{ii}]_{JJ}^{-1} [\mathbf{H}_{ii}]_{JJ} [\mathbf{M}_{ii}]_{JJ}^{-1}, \quad (6.14)$$

que expressa o deslocamento no limite de confiança conjunto, para os parâmetros  $\boldsymbol{\xi}$ .

Por fim, a variação percentual da estimativa de cada parâmetro  $\xi_j \in \boldsymbol{\xi}$ , quando retirada a  $i$ -ésima observação da amostra, é dada por,

$$RC(\hat{\xi}_j) = \left| \frac{\hat{\xi}_j^{(i)} - \hat{\xi}_j}{\hat{\xi}_j} \right| \times 100, \quad (6.15)$$



em que  $\hat{\xi}_{j(i)}$  é a estimativa de  $\xi_j \in \boldsymbol{\xi}$  sem a  $i$ -ésima observação, e essa medida é calculada para as observações acusadas como discrepantes pelos gráficos de diagnóstico das medidas anteriormente explicadas.

### 6.1.1 Modelo Independente

Seja a matriz,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_{60} \end{pmatrix} \otimes \mathbf{I}_3, \quad (6.16)$$

em que  $\mathbf{A}_1 = \dots = \mathbf{A}_{60} = \mathbf{I}_3$ .

O modelo independente possui a seguinte matriz de especificação,

$$\mathbf{X} = \begin{pmatrix} & | & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \mathbf{A} & | & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ & | & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{pmatrix}, \quad (6.17)$$

em que  $\mathbf{X}_p = (X_{p,1}, X_{p,1}, X_{p,1}, \dots, X_{p,60}, X_{p,60}, X_{p,60})'$  representa cada covariável, para todo  $p = 1, 2, 3$ . As covariáveis se repetem no vetor, para permitir a modelagem de cada um dos três logitos, para cada um dos códons.

O conjunto de parâmetros do modelo é dado por,

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix}, \quad (6.18)$$

tal que,  $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})'$  e  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ .

### 6.1.2 Modelo Igualmente Preditivo

O modelo igualmente preditivo possui a seguinte matriz de especificação,

$$\mathbf{X} = \begin{pmatrix} & | & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \mathbf{A} & | & \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{Z}_{13} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ & | & \mathbf{Z}_{11} + \mathbf{Z}_{21} & \mathbf{Z}_{12} + \mathbf{Z}_{22} & \mathbf{Z}_{13} + \mathbf{Z}_{23} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{pmatrix}, \quad (6.19)$$

em que  $\mathbf{X}_p = (X_{p,1}, X_{p,1}, X_{p,1}, \dots, X_{p,60}, X_{p,60}, X_{p,60})'$ , para todo  $p = 1, 2, 3$ ,  $\mathbf{Z}_{kj} = (Z_{kj,1}, Z_{kj,1}, Z_{kj,1}, \dots, Z_{kj,60}, Z_{kj,60}, Z_{kj,60})'$  para todo  $k, j = 1, 2, 3$  e  $\mathbf{A}$  é conforme definida no modelo independente. Assim como no modelo independente, as variáveis  $Z_{kj,i}$  se repetem no vetor, para permitir a modelagem de cada um dos três logitos, para cada um dos códons.

O conjunto de parâmetros do modelo é dado por,

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}, \quad (6.20)$$

tal que,  $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})'$ ,  $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)'$  e  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ .

### 6.1.3 Estrutura Markoviana de Primeira Ordem

A estrutura markoviana de primeira ordem possui a seguinte matriz de especificação,

$$\mathbf{X} = \begin{pmatrix} & | & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \mathbf{A} & | & \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{Z}_{13} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ & | & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{21} & \mathbf{Z}_{22} & \mathbf{Z}_{23} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{pmatrix}, \quad (6.21)$$

em que  $\mathbf{X}_p = (X_{p,1}, X_{p,1}, X_{p,1}, \dots, X_{p,60}, X_{p,60}, X_{p,60})'$ , para todo  $p = 1, 2, 3$ ,  $\mathbf{Z}_{kj} = (Z_{kj,1}, Z_{kj,1}, Z_{kj,1}, \dots, Z_{kj,60}, Z_{kj,60}, Z_{kj,60})'$  para todo  $k, j = 1, 2, 3$  e  $\mathbf{A}$  é conforme definida no modelo independente.

O conjunto de parâmetros do modelo é dado por,

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}, \quad (6.22)$$

tal que,  $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})'$ ,  $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23})'$  e  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ .

### 6.1.4 Modelo Aditivo

O modelo aditivo possui a seguinte matriz de especificação,

$$\mathbf{X} = \begin{pmatrix} | & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \mathbf{A} & | & \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{Z}_{13} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ | & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{Z}_{13} & \mathbf{Z}_{21} & \mathbf{Z}_{22} & \mathbf{Z}_{23} & \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{pmatrix}, (6.23)$$

em que  $\mathbf{X}_p = (X_{p,1}, X_{p,1}, X_{p,1}, \dots, X_{p,60}, X_{p,60}, X_{p,60})'$ , para todo  $p = 1, 2, 3$ ,  $\mathbf{Z}_{kj} = (Z_{kj,1}, Z_{kj,1}, Z_{kj,1}, \dots, Z_{kj,60}, Z_{kj,60}, Z_{kj,60})'$  para todo  $k, j = 1, 2, 3$  e  $\mathbf{A}$  é conforme definida no modelo independente.

O conjunto de parâmetros do modelo é dado por,

$$\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}, (6.24)$$

tal que,  $\boldsymbol{\alpha} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})'$ ,  $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \gamma_{33})'$  e  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ .

## 6.2 Resultados

As Figuras 6.1, 6.2, 6.3 e 6.4 apresentam os gráficos de diagnóstico obtidos para os quatro modelos logísticos regressivos, para cada uma das três posições dos códons. Em geral os códons detectados como pontos de influência são comuns aos quatro modelos. A estrutura markoviana de primeira ordem e o modelo aditivo apresentam menos pontos de influência do que o modelo independente e igualmente preditivo. Além disso a segunda posição do códon é a que mais apresenta pontos de influência, quando levada em consideração a distância de Cook.

As Tabelas 6.1, 6.2, 6.3 e 6.4 apresentam as estimativas dos parâmetros quando removido cada códon apontado pelos gráficos de diagnóstico. Também são apresentadas as variações percentuais das estimativas dos parâmetros quando cada códon é retirado

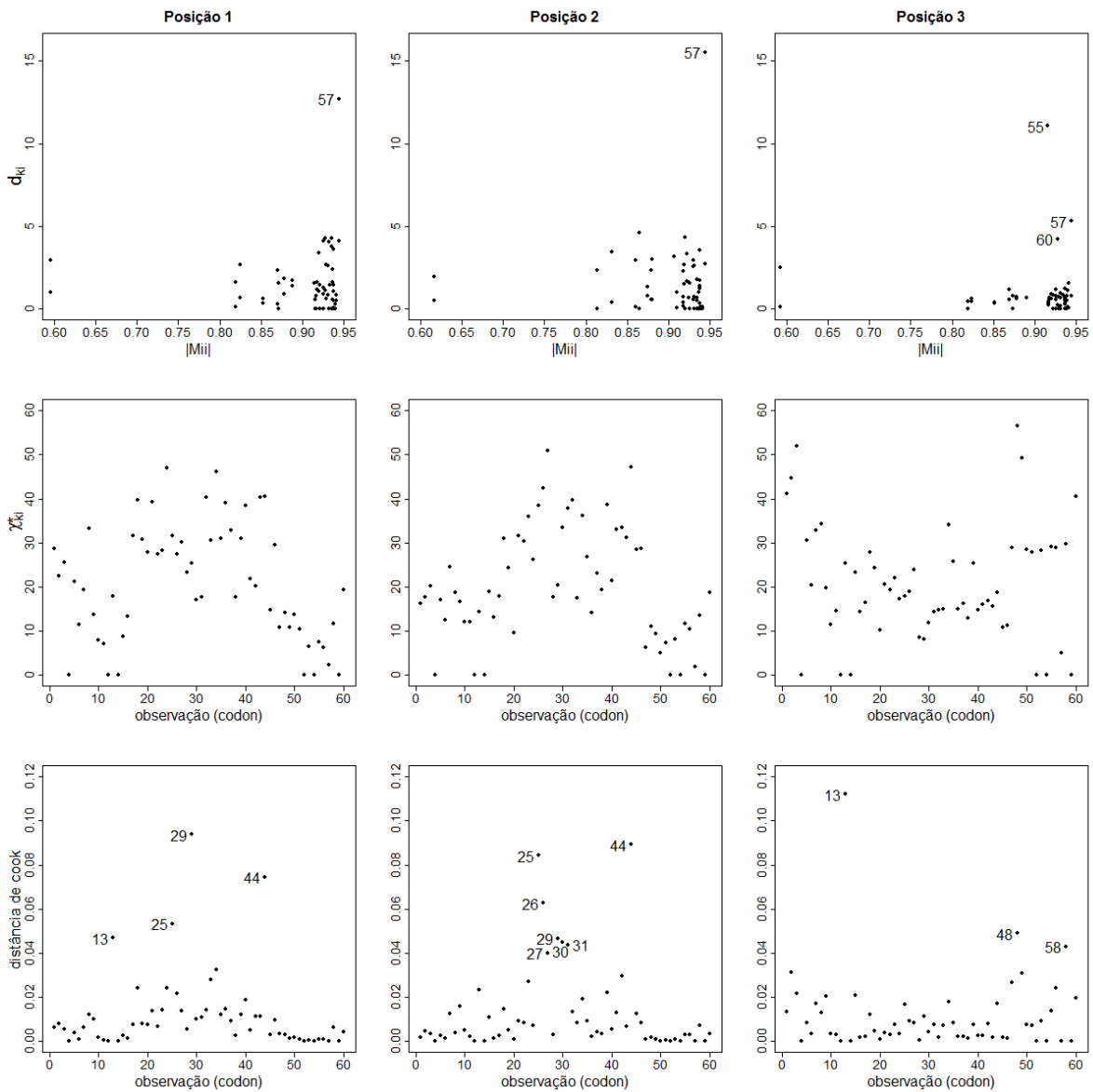


Figura 6.1: Diagnóstico do Modelo Independente

da amostra. Vale ressaltar que os códons nunca são retirados simultaneamente para essas análises, mas sempre individualmente.

No modelo independente, os códons 25 (TAC) e 26 (CAC) são os que possuem mais impacto nas estimativas dos parâmetros, quando retirados da amostra, pois as estimativas de quase todos os parâmetros sofrem grandes alterações. Além disso, os parâmetros  $\alpha_{31}$ ,  $\alpha_{32}$ ,  $\beta_1$  e  $\beta_3$  são os que sofrem mais alterações com as retiradas de

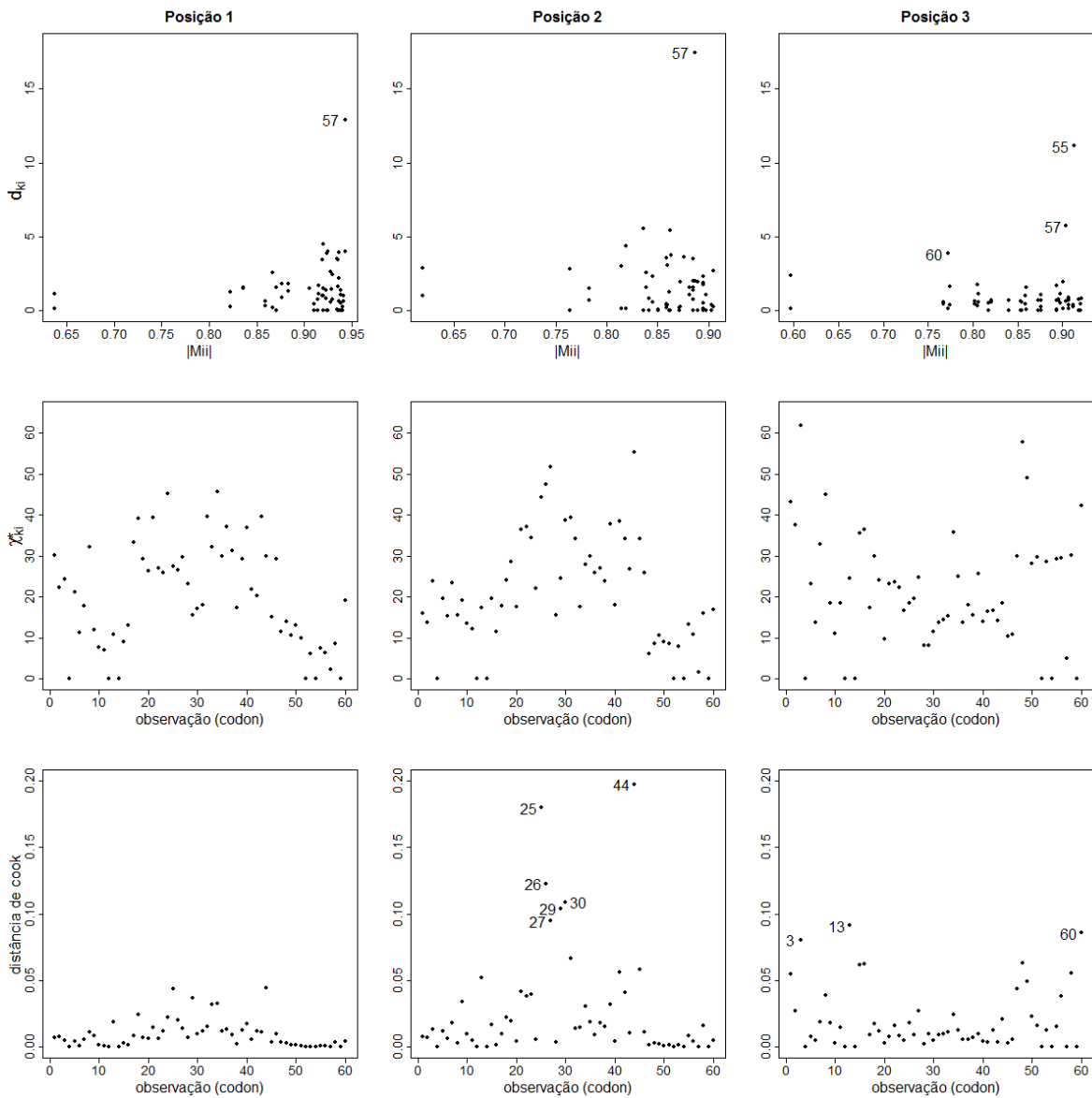


Figura 6.2: Diagnóstico do Modelo Iguamente Preditivo

códons da amostra.

Já no modelo igualmente preditivo, a maior impacto nas estimativas dos parâmetros  $\tilde{\mathbf{A}}_{\odot}$  dos códons 3 (ATT), 26 (CAC) e 27 (AAC). Os parâmetros que sofrem mais alterações com as retiradas de códons da amostra são  $\alpha_{11}$ ,  $\alpha_{12}$ ,  $\alpha_{13}$ ,  $\alpha_{31}$  e  $\alpha_{32}$ .

A estrutura Markoviana de primeira ordem não sofre tanta influência com a retirada de códons da amostra. Apenas o códon 13 (TGT) têm mais impacto na estimativa de

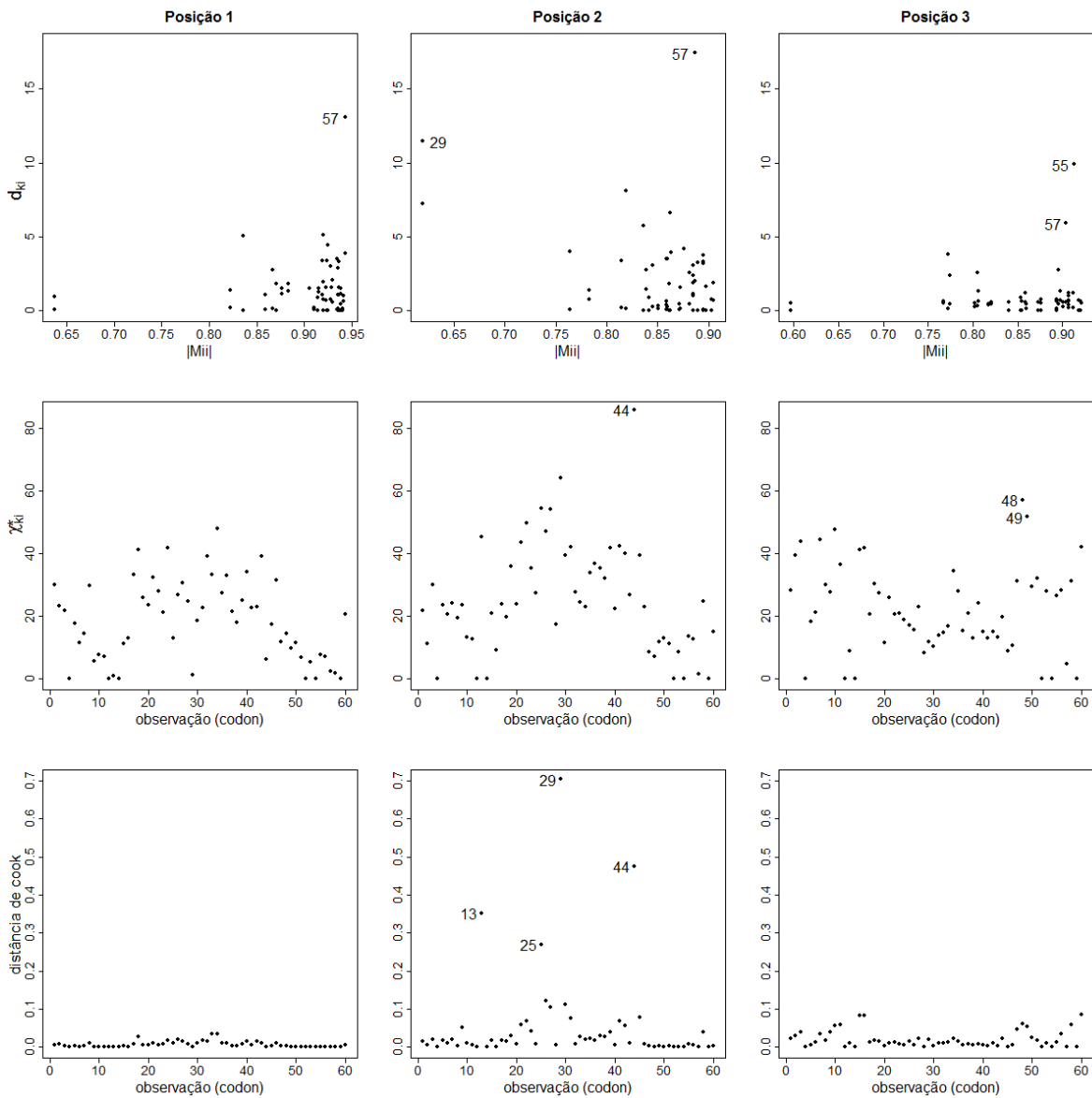


Figura 6.3: Diagnóstico da Estrutura Markoviana de Primeira Ordem

quase todos os parâmetros, e apenas os parâmetros  $\alpha_{33}$  e  $\beta_2$  sofrem mais alterações com as retiradas de códons da amostra.

Por fim no modelo aditivo quase todos os parâmetros sofrem alterações com a retirada dos códons 13 (TGT), 25 (TAC), 29 (TGC) e 44 (TGA).

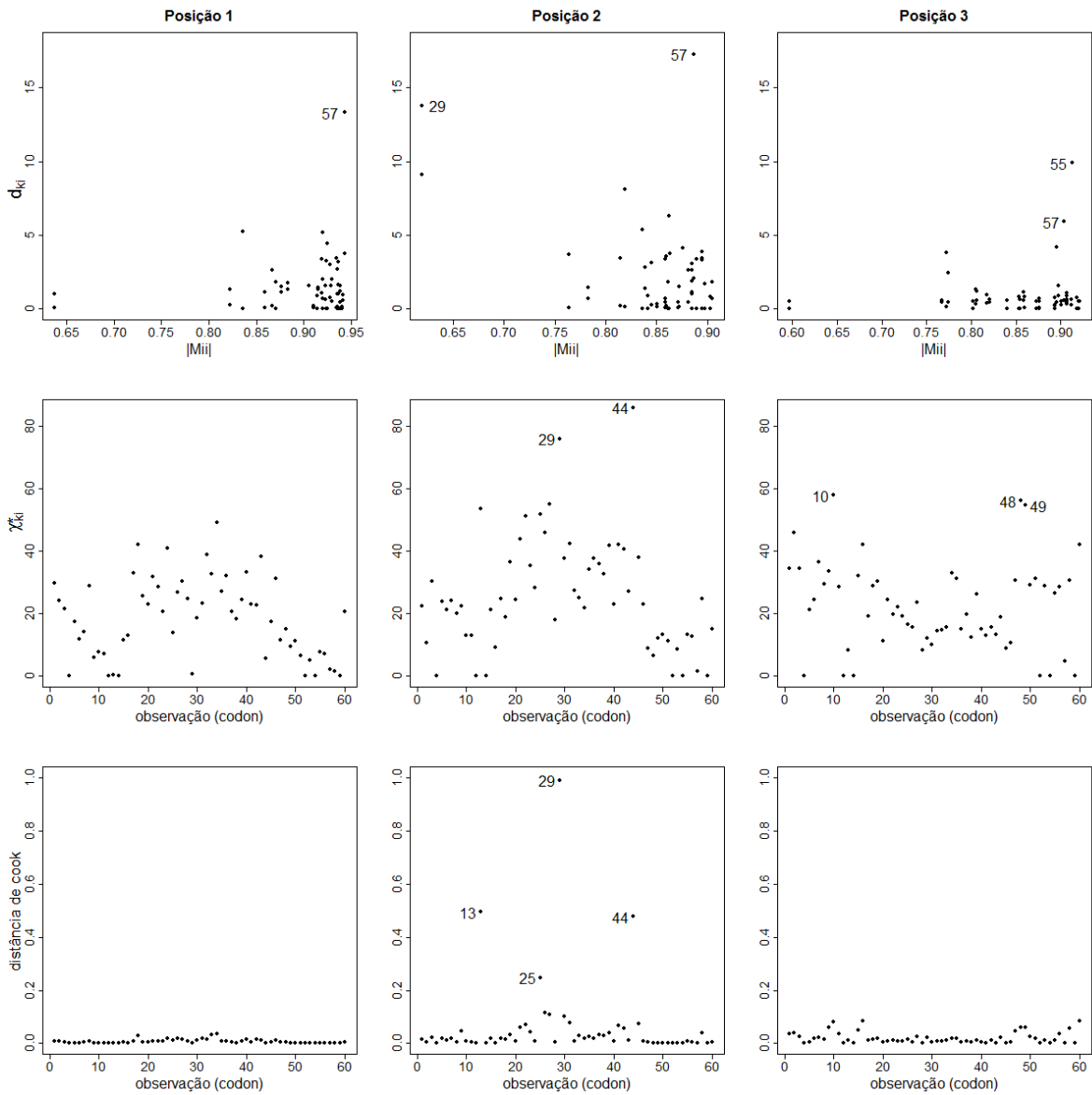


Figura 6.4: Diagnóstico do Modelo Aditivo

Tabela 6.1: Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Independente

Obs. Retirada	$\hat{\alpha}_{11(i)}$	$RC(\hat{\alpha}_{11(i)})$	$\hat{\alpha}_{12(i)}$	$RC(\hat{\alpha}_{12(i)})$	$\hat{\alpha}_{13(i)}$	$RC(\hat{\alpha}_{13(i)})$	$\hat{\alpha}_{21(i)}$	$RC(\hat{\alpha}_{21(i)})$	$\hat{\alpha}_{22(i)}$	$RC(\hat{\alpha}_{22(i)})$	$\hat{\alpha}_{23(i)}$	$RC(\hat{\alpha}_{23(i)})$
13 (TGT)	-1,4208	39,90	-1,3044	45,72	-2,2100	23,17	-2,0807	21,67	-2,4934	18,50	-2,8658	15,90
25 (TAC)	-0,3324	67,27	-0,1937	78,37	-1,1248	39,77	-1,1398	33,35	-1,6963	19,38	-1,9209	22,31
26 (CAC)	-2,2205	118,64	-2,0459	128,55	-3,0790	64,89	-2,7915	63,24	-3,3704	60,18	-3,5595	43,95
27 (AAC)	-0,8542	15,89	-0,9194	2,70	-1,7200	7,89	-1,6650	2,64	-2,3281	10,65	-2,4415	1,26
29 (TGC)	-1,2594	24,00	-1,1409	27,45	-2,1273	13,92	-1,9432	13,63	-2,3511	11,74	-2,7464	11,07
30 (CGC)	-1,2617	24,24	-1,1259	25,78	-2,1029	12,62	-1,9378	13,31	-2,3273	10,61	-2,7969	13,11
31 (AGC)	-0,4319	57,47	-0,3417	61,83	-1,2369	33,76	-1,1715	31,50	-1,5366	26,98	-2,0831	15,75
44 (TGA)	-0,8966	11,72	-0,7934	11,37	-1,7231	7,72	-1,6313	4,61	-2,0877	0,78	-2,7231	10,13
48 (CTG)	-0,9817	3,33	-0,8303	7,25	-1,7990	3,66	-1,6260	4,92	-2,0221	3,90	-2,3900	3,35
55 (CAG)	-1,0844	6,77	-0,9570	6,90	-1,9340	3,57	-1,7697	3,48	-2,1774	3,48	-2,5312	2,37
57 (GAG)	-1,0160	0,04	-0,8955	0,04	-1,8682	0,05	-1,7106	0,03	-2,1050	0,04	-2,4731	0,02
58 (TGG)	-0,9889	2,63	-0,8694	2,88	-1,8375	1,60	-1,6891	1,23	-2,0869	0,82	-2,4753	0,11
60 (GGG)	-0,9282	8,61	-0,8069	9,86	-1,8044	3,37	-1,6303	4,67	-2,0216	3,92	-2,4335	1,58
Obs. Retirada	$\hat{\alpha}_{31(i)}$	$RC(\hat{\alpha}_{31(i)})$	$\hat{\alpha}_{32(i)}$	$RC(\hat{\alpha}_{32(i)})$	$\hat{\alpha}_{33(i)}$	$RC(\hat{\alpha}_{33(i)})$	$\hat{\beta}_{1(i)}$	$RC(\hat{\beta}_{1(i)})$	$\hat{\beta}_{2(i)}$	$RC(\hat{\beta}_{2(i)})$	$\hat{\beta}_{3(i)}$	$RC(\hat{\beta}_{3(i)})$
13 (TGT)	-0,5523	218,20	-0,5918	178,91	-2,9856	14,51	-0,0057	7,60	0,2210	22,71	0,0873	4,39
25 (TAC)	0,3226	285,86	0,3271	254,16	-2,0717	20,54	-0,0012	80,89	0,0879	51,21	0,1267	51,63
26 (CAC)	-1,3084	653,75	-1,2816	504,06	-3,6731	40,88	-0,0132	113,15	0,3383	87,85	0,1687	101,82
27 (AAC)	-0,1293	25,54	-0,0613	71,11	-2,4560	5,81	-0,0159	157,42	0,2186	21,37	0,2242	168,22
29 (TGC)	-0,4346	150,34	-0,4652	119,27	-2,8597	9,68	-0,0056	9,31	0,2047	13,65	0,0833	0,28
30 (CGC)	-0,3975	129,01	-0,4147	95,46	-2,8083	7,71	-0,0096	55,49	0,2226	23,58	0,1142	36,58
31 (AGC)	0,3257	287,65	0,3261	253,70	-2,0714	20,56	-0,0039	37,36	0,1084	39,83	-0,0083	109,92
44 (TGA)	-0,1346	22,49	-0,2752	29,70	-2,6016	0,22	-0,0003	95,95	0,1385	23,12	0,2254	169,69
48 (CTG)	-0,1147	33,95	-0,1520	28,36	-2,8291	8,51	-0,0055	11,32	0,1693	5,98	0,0807	3,46
55 (CAG)	-0,2355	35,69	-0,2736	28,96	-2,7330	4,82	-0,0063	2,32	0,1879	4,31	0,0789	5,62
57 (GAG)	-0,1740	0,26	-0,2126	0,21	-2,6099	0,10	-0,0062	0,04	0,1802	0,04	0,0834	0,18
58 (TGG)	-0,1553	10,53	-0,1960	7,60	-2,6558	1,86	-0,0058	6,69	0,1754	2,60	0,0918	9,85
60 (GGG)	-0,0918	47,09	-0,1316	37,98	-2,6674	2,30	-0,0060	3,77	0,1699	5,66	0,0764	8,59



Tabela 6.2: Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Igualmente Preditivo

Obs. Retirada	$\hat{\alpha}_{11(i)}$	$RC(\hat{\alpha}_{11(i)})$	$\hat{\alpha}_{12(i)}$	$RC(\hat{\alpha}_{12(i)})$	$\hat{\alpha}_{13(i)}$	$RC(\hat{\alpha}_{13(i)})$	$\hat{\alpha}_{21(i)}$	$RC(\hat{\alpha}_{21(i)})$	$\hat{\alpha}_{22(i)}$	$RC(\hat{\alpha}_{22(i)})$	$\hat{\alpha}_{23(i)}$	$RC(\hat{\alpha}_{23(i)})$
3 (ATT)	3,0419	4243,97	3,2057	3736,57	2,4712	384,09	2,3275	343,33	2,1811	270,64	2,0067	222,77
13 (TGT)	-0,8746	1091,51	-0,7260	968,84	-1,7255	98,36	-1,7296	80,83	-2,0849	63,11	-2,4359	49,02
25 (TAC)	0,6308	959,36	0,8003	857,86	-0,1230	85,87	-0,4958	48,17	-0,9616	24,77	-1,1515	29,56
26 (CAC)	-1,9097	2501,66	-1,6945	2127,97	-2,7565	216,89	-2,6799	180,18	-3,1905	149,61	-3,3531	105,13
27 (AAC)	3,1872	4441,93	3,1688	3692,44	2,6232	401,56	2,5612	367,77	2,0446	259,96	2,0785	227,15
29 (TGC)	-0,5590	661,47	-0,4079	588,22	-1,3897	59,76	-1,4534	51,95	-1,7966	40,56	-2,1743	33,02
30 (CGC)	-0,3393	362,21	-0,1646	297,02	-1,1263	29,49	-1,2034	25,82	-1,5091	18,07	-1,9657	20,26
44 (TGA)	-0,0336	54,25	0,0984	17,80	-0,8175	6,02	-1,0405	8,79	-1,4241	11,41	-2,0437	25,02
55 (CAG)	-0,2740	273,26	-0,1109	232,71	-1,0782	23,95	-1,1482	20,05	-1,4863	16,28	-1,8246	11,62
57 (GAG)	-0,0763	3,98	0,0806	3,51	-0,8735	0,42	-0,9593	0,30	-1,2815	0,26	-1,6374	0,17
60 (GGG)	0,0166	122,55	0,1728	106,79	-0,8037	7,61	-0,8567	10,43	-1,1760	8,00	-1,5763	3,57
Obs. Retirada	$\hat{\alpha}_{31(i)}$	$RC(\hat{\alpha}_{31(i)})$	$\hat{\alpha}_{32(i)}$	$RC(\hat{\alpha}_{32(i)})$	$\hat{\alpha}_{33(i)}$	$RC(\hat{\alpha}_{33(i)})$	$\hat{\gamma}_{1(i)}$	$RC(\hat{\gamma}_{1(i)})$	$\hat{\gamma}_{2(i)}$	$RC(\hat{\gamma}_{2(i)})$	$\hat{\gamma}_{3(i)}$	$RC(\hat{\gamma}_{3(i)})$
3 (ATT)	4,7307	577,95	4,7121	592,47	2,3045	234,19	-1,0937	163,61	0,1553	16,03	0,5902	28,63
13 (TGT)	-0,1346	119,29	-0,1525	122,42	-2,5477	48,35	-0,3613	12,91	0,2405	30,08	0,8523	3,08
25 (TAC)	1,0651	52,63	1,0921	60,50	-1,3081	23,83	-0,3064	26,15	0,2503	35,38	0,9792	18,42
26 (CAC)	-1,1162	259,97	-1,0615	256,00	-3,4539	101,12	-0,4373	5,41	0,3228	74,59	0,8326	0,69
27 (AAC)	4,6402	564,98	4,6614	585,03	2,2539	231,24	-0,7979	92,31	-0,6856	470,83	0,7921	4,20
29 (TGC)	0,1330	80,94	0,1241	81,76	-2,2721	32,30	-0,3587	13,54	0,2396	29,57	0,8252	0,20
30 (CGC)	0,4794	31,30	0,4865	28,51	-1,9097	11,20	-0,4564	10,00	0,2129	15,14	0,8077	2,32
44 (TGA)	0,5416	22,39	0,4220	37,99	-1,9063	11,00	-0,2792	32,70	0,2193	18,63	0,8676	4,93
55 (CAG)	0,4920	29,50	0,4763	30,00	-1,9852	15,60	-0,4097	1,25	0,2043	10,49	0,8254	0,18
57 (GAG)	0,6947	0,45	0,6774	0,46	-1,7225	0,30	-0,4146	0,07	0,1852	0,16	0,8264	0,06
60 (GGG)	0,8073	15,69	0,7882	15,83	-1,7503	1,92	-0,4223	1,79	0,1725	6,73	0,7944	3,93
Obs. Retirada	$\hat{\beta}_{1(i)}$	$RC(\hat{\beta}_{1(i)})$	$\hat{\beta}_{2(i)}$	$RC(\hat{\beta}_{2(i)})$	$\hat{\beta}_{3(i)}$	$RC(\hat{\beta}_{3(i)})$						
3 (ATT)	-0,0141	39,77	-0,2054	284,03	-0,2183	429,58						
13 (TGT)	-0,0098	2,62	0,1969	76,44	-0,0406	1,61						
25 (TAC)	-0,0082	18,32	0,0384	65,55	-0,0362	12,08						
26 (CAC)	-0,0192	90,49	0,3526	215,96	0,0121	129,28						
27 (AAC)	-0,0158	56,03	-0,2326	308,44	0,2421	687,34						
29 (TGC)	-0,0099	1,62	0,1652	48,01	-0,0437	6,14						
30 (CGC)	-0,0145	43,72	0,1621	45,26	-0,0236	42,70						
44 (TGA)	-0,0052	49,00	0,0861	22,83	0,0743	280,16						
55 (CAG)	-0,0105	3,54	0,1352	21,13	-0,0508	23,33						
57 (GAG)	-0,0101	0,03	0,1119	0,31	-0,0413	0,28						
60 (GGG)	-0,0098	3,01	0,0999	10,48	-0,0409	0,66						

Tabela 6.3: Estimativas dos Parâmetros Retirando Observações Discrepantes da Estrutura Markoviana

Obs. Retirada	$\hat{\alpha}_{11(i)}$	$RC(\hat{\alpha}_{11(i)})$	$\hat{\alpha}_{12(i)}$	$RC(\hat{\alpha}_{12(i)})$	$\hat{\alpha}_{13(i)}$	$RC(\hat{\alpha}_{13(i)})$	$\hat{\alpha}_{21(i)}$	$RC(\hat{\alpha}_{21(i)})$	$\hat{\alpha}_{22(i)}$	$RC(\hat{\alpha}_{22(i)})$	$\hat{\alpha}_{23(i)}$	$RC(\hat{\alpha}_{23(i)})$
13 (TGT)	-0,6146	120,47	-0,4192	112,62	-1,4498	160,33	-1,1048	146,25	-1,3881	157,79	-1,7478	174,58
25 (TAC)	2,0702	31,06	2,3894	28,06	1,3398	44,25	1,0882	54,45	1,1017	54,13	1,2409	47,05
29 (TGC)	4,1628	38,63	4,5088	35,75	3,6639	52,47	3,6275	51,85	3,6300	51,12	3,6256	54,71
44 (TGA)	2,5912	13,71	2,9370	11,58	1,9198	20,11	1,7423	27,07	2,0213	15,85	1,8658	20,38
48 (CTG)	3,0533	1,68	3,3941	2,18	2,4894	3,59	2,4766	3,67	2,4705	2,85	2,3986	2,35
49 (ATG)	3,0278	0,83	3,3247	0,10	2,4310	1,16	2,4213	1,36	2,4203	0,76	2,3552	0,50
55 (CAG)	2,9559	1,57	3,2821	1,19	2,3596	1,81	2,3471	1,75	2,3514	2,11	2,3039	1,69
57 (GAG)	3,0024	0,01	3,3211	0,01	2,4021	0,04	2,3885	0,02	2,4014	0,03	2,3434	0,00
Obs. Retirada	$\hat{\alpha}_{31(i)}$	$RC(\hat{\alpha}_{31(i)})$	$\hat{\alpha}_{32(i)}$	$RC(\hat{\alpha}_{32(i)})$	$\hat{\alpha}_{33(i)}$	$RC(\hat{\alpha}_{33(i)})$	$\hat{\gamma}_{11(i)}$	$RC(\hat{\gamma}_{11(i)})$	$\hat{\gamma}_{12(i)}$	$RC(\hat{\gamma}_{12(i)})$	$\hat{\gamma}_{13(i)}$	$RC(\hat{\gamma}_{13(i)})$
13 (TGT)	-0,3949	112,53	-0,3766	111,86	-2,7728	460,03	-1,1367	23,68	-0,1552	27,43	0,6627	43,36
25 (TAC)	1,9810	37,15	2,0959	34,00	-0,3035	139,40	-1,1886	20,20	0,1763	182,44	0,6569	42,10
29 (TGC)	4,3515	38,07	4,3620	37,37	1,9525	153,51	-1,6631	11,66	-0,3509	64,08	0,3745	18,99
44 (TGA)	2,5834	18,03	2,5791	18,78	0,2493	67,63	-1,3958	6,29	0,0492	123,01	0,5100	10,32
48 (CTG)	3,2282	2,43	3,2473	2,26	0,5601	27,28	-1,4529	2,45	-0,2352	9,96	0,4672	1,06
49 (ATG)	3,1723	0,65	3,1927	0,54	0,5800	24,69	-1,4908	0,09	-0,1848	13,62	0,4670	1,03
55 (CAG)	3,1094	1,34	3,1347	1,28	0,6652	13,63	-1,4994	0,67	-0,2055	3,90	0,4656	0,72
57 (GAG)	3,1512	0,02	3,1750	0,01	0,7677	0,33	-1,4894	0,00	-0,2137	0,08	0,4617	0,14
Obs. Retirada	$\hat{\gamma}_{21(i)}$	$RC(\hat{\gamma}_{21(i)})$	$\hat{\gamma}_{22(i)}$	$RC(\hat{\gamma}_{22(i)})$	$\hat{\gamma}_{23(i)}$	$RC(\hat{\gamma}_{23(i)})$	$\hat{\beta}_{1(i)}$	$RC(\hat{\beta}_{1(i)})$	$\hat{\beta}_{2(i)}$	$RC(\hat{\beta}_{2(i)})$	$\hat{\beta}_{3(i)}$	$RC(\hat{\beta}_{3(i)})$
13 (TGT)	0,2664	26,70	2,0657	22,52	1,6948	42,96	-0,0115	53,94	0,1871	241,70	-0,2354	49,39
25 (TAC)	0,3469	65,00	2,7526	3,24	3,4400	15,78	-0,0384	54,64	0,0639	148,41	-0,6543	40,69
29 (TGC)	0,1569	25,35	2,5000	6,24	2,6346	11,33	-0,0209	15,89	-0,2839	115,06	-0,4135	11,08
44 (TGA)	0,3821	81,74	3,4865	30,76	2,8126	5,34	-0,0438	76,28	0,0329	124,94	-0,7488	61,02
48 (CTG)	0,2175	3,47	2,6063	2,25	2,9273	1,48	-0,0232	6,88	-0,1505	13,99	-0,4404	5,30
49 (ATG)	0,2232	6,15	2,6408	0,96	2,9537	0,59	-0,0243	2,34	-0,1384	4,81	-0,4463	4,03
55 (CAG)	0,2139	1,75	2,6725	0,23	2,9779	0,23	-0,0251	0,78	-0,1262	4,39	-0,4714	1,35
57 (GAG)	0,2103	0,05	2,6664	0,01	2,9716	0,02	-0,0248	0,03	-0,1319	0,08	-0,4653	0,05

Tabela 6.4: Estimativas dos Parâmetros Retirando Observações Discrepantes do Modelo Aditivo

Obs. Retirada	$\hat{\alpha}_{11(i)}$	$RC(\hat{\alpha}_{11(i)})$	$\hat{\alpha}_{12(i)}$	$RC(\hat{\alpha}_{12(i)})$	$\hat{\alpha}_{13(i)}$	$RC(\hat{\alpha}_{13(i)})$	$\hat{\alpha}_{21(i)}$	$RC(\hat{\alpha}_{21(i)})$	$\hat{\alpha}_{22(i)}$	$RC(\hat{\alpha}_{22(i)})$	$\hat{\alpha}_{23(i)}$	$RC(\hat{\alpha}_{23(i)})$
10 (CAT)	3,7623	2,16	4,1422	2,63	3,2386	2,99	3,2452	3,37	3,2895	3,88	3,3405	5,31
13 (TGT)	-0,2958	108,03	-0,0740	101,83	-1,0702	134,03	-0,4626	114,73	-0,8067	125,48	-1,1420	136,00
25 (TAC)	1,0821	70,62	1,3061	67,64	0,4143	86,83	0,9067	71,12	0,4016	87,32	0,1900	94,01
29 (TGC)	-0,2264	106,15	-0,0065	100,16	-0,9943	131,62	-0,3829	112,20	-0,7289	123,02	-1,0860	134,23
44 (TGA)	0,1195	96,76	0,2803	93,06	-0,6133	119,50	-0,0087	100,28	-0,4126	113,03	-1,0299	132,47
48 (CTG)	3,7620	2,15	4,1330	2,40	3,2630	3,77	3,2608	3,87	3,2600	2,95	3,2395	2,12
49 (ATG)	3,8025	3,25	4,1317	2,37	3,2768	4,21	3,2782	4,42	3,2767	3,47	3,2632	2,87
55 (CAG)	3,6296	1,44	3,9913	1,11	3,0940	1,61	3,0908	1,54	3,1119	1,73	3,1319	1,27
57 (GAG)	3,6812	0,04	4,0346	0,03	3,1424	0,07	3,1378	0,05	3,1651	0,05	3,1714	0,02
Obs. Retirada	$\hat{\alpha}_{31(i)}$	$RC(\hat{\alpha}_{31(i)})$	$\hat{\alpha}_{32(i)}$	$RC(\hat{\alpha}_{32(i)})$	$\hat{\alpha}_{33(i)}$	$RC(\hat{\alpha}_{33(i)})$	$\hat{\gamma}_{11(i)}$	$RC(\hat{\gamma}_{11(i)})$	$\hat{\gamma}_{12(i)}$	$RC(\hat{\gamma}_{12(i)})$	$\hat{\gamma}_{13(i)}$	$RC(\hat{\gamma}_{13(i)})$
10 (CAT)	4,2533	1,21	4,2848	1,42	1,8771	3,32	-1,6748	3,26	-0,2999	0,61	0,3340	10,19
13 (TGT)	-0,5301	112,61	-0,5175	112,25	-2,9172	260,57	-1,3667	15,73	-0,3998	34,12	0,4845	30,30
25 (TAC)	0,6879	83,63	0,7241	82,86	-1,6808	192,51	-1,4755	9,03	-0,5203	74,56	0,4964	33,51
29 (TGC)	-0,5329	112,68	-0,5115	112,11	-2,9117	260,26	-1,3788	14,99	-0,4173	39,99	0,4792	28,86
44 (TGA)	-0,1333	103,17	-0,2442	105,78	-2,5763	241,81	-1,3807	14,87	-0,5207	74,70	0,5393	45,03
48 (CTG)	4,2976	2,27	4,3109	2,04	1,6213	10,76	-1,5868	2,16	-0,3305	10,87	0,3795	2,07
49 (ATG)	4,3030	2,40	4,3167	2,18	1,7013	6,36	-1,6353	0,83	-0,2888	3,10	0,3757	1,03
55 (CAG)	4,1533	1,17	4,1778	1,11	1,7059	6,11	-1,6320	0,63	-0,2883	3,29	0,3733	0,40
57 (GAG)	4,2010	0,03	4,2234	0,03	1,8136	0,18	-1,6218	0,01	-0,2977	0,13	0,3712	0,19
Obs. Retirada	$\hat{\gamma}_{21(i)}$	$RC(\hat{\gamma}_{21(i)})$	$\hat{\gamma}_{22(i)}$	$RC(\hat{\gamma}_{22(i)})$	$\hat{\gamma}_{23(i)}$	$RC(\hat{\gamma}_{23(i)})$	$\hat{\gamma}_{31(i)}$	$RC(\hat{\gamma}_{31(i)})$	$\hat{\gamma}_{32(i)}$	$RC(\hat{\gamma}_{32(i)})$	$\hat{\gamma}_{33(i)}$	$RC(\hat{\gamma}_{33(i)})$
10 (CAT)	0,0176	124,02	-0,7930	1,37	-0,3048	6,25	0,1738	20,48	3,0132	14,88	3,1389	4,53
13 (TGT)	1,0963	1597,98	0,3569	145,62	0,5851	304,01	0,3879	168,97	1,9437	25,89	1,6657	44,53
25 (TAC)	1,3643	1964,24	0,4394	156,17	0,8744	404,88	0,2859	98,20	1,2704	51,57	1,3839	53,91
29 (TGC)	1,1794	1711,59	0,4387	156,08	0,6958	342,60	0,3867	168,11	1,9328	26,31	1,3346	55,56
44 (TGA)	1,3408	1931,99	0,4929	163,00	0,9279	423,52	0,2854	97,84	1,4196	45,88	0,6827	77,27
48 (CTG)	-0,0586	19,99	-0,7588	3,01	-0,2399	16,35	0,1399	3,00	2,5320	3,46	2,9262	2,55
49 (ATG)	-0,0195	73,37	-0,7997	2,22	-0,2497	12,96	0,1578	9,38	2,5673	2,12	2,9639	1,29
55 (CAG)	-0,0769	5,04	-0,7782	0,52	-0,2913	1,58	0,1499	3,95	2,6369	0,54	3,0169	0,47
57 (GAG)	-0,0737	0,70	-0,7824	0,02	-0,2875	0,23	0,1445	0,17	2,6237	0,03	3,0038	0,04
Obs. Retirada	$\hat{\beta}_{1(i)}$	$RC(\hat{\beta}_{1(i)})$	$\hat{\beta}_{2(i)}$	$RC(\hat{\beta}_{2(i)})$	$\hat{\beta}_{3(i)}$	$RC(\hat{\beta}_{3(i)})$						
10 (CAT)	-0,0243	5,15	-0,2210	1,24	-0,5189	8,83						
13 (TGT)	0,0017	107,48	0,0639	129,29	-0,1901	60,12						
25 (TAC)	0,0145	162,92	-0,1580	27,64	-0,0714	85,03						
29 (TGC)	0,0027	111,60	0,0509	123,31	-0,1847	61,25						
44 (TGA)	0,0108	146,68	-0,0371	83,01	0,0394	108,26						
48 (CTG)	-0,0205	11,44	-0,2453	12,38	-0,4378	8,17						
49 (ATG)	-0,0209	9,79	-0,2447	12,10	-0,4409	7,51						
55 (CAG)	-0,0235	1,85	-0,2105	3,57	-0,4859	1,93						
57 (GAG)	-0,0231	0,13	-0,2180	0,16	-0,4772	0,10						

# Capítulo 7

## Considerações Finais

A partir da aplicação dos modelos propostos, uma primeira constatação feita é de que ao modelarmos as bases nitrogenadas como dados binários, classificando-as como *purinas* ou *pirimidinas*, há uma grande perda de informação, quando comparados aos modelos com respostas multinomiais, em que as bases são modeladas como Timina, Citosina, Adenina e Guanina. As Figuras 5.1, 5.2, 5.3 e 5.4, evidenciam essa perda de informação, uma vez que é visível nos gráficos um melhor ajuste pelos modelos multinomiais.

Em uma primeira análise dos resultados obtidos nos modelos multinomiais, não apenas os valores do AIC, BIC e SQE indicam que o modelo aditivo é mais adequado para os dados de DNA, como também a validação cruzada é consistente com esses resultados, apresentando menor QME e menor variação da SQE.

Há no entanto, alguns pontos interessantes a serem notados sobre esses modelos. O primeiro é que os valores do AIC e BIC do modelo de semi-locação e transição equivalem a 1,0044 e 1,0050 vezes os valores do modelo aditivo, respectivamente. Portanto, proporcionalmente, os valores dessas medidas, obtidos para o melhor modelo dentre os baseados na representação de Bahadur não são muito maiores do que os obtidos para o melhor modelo regressivo.

Outro ponto importante no estudo de modelos para seqüências de DNA, é o de que a suposição de independência entre as três bases que compõem um códon não é

verdadeira, como mostraram as medidas de ajuste para os modelos sob essa suposição. Os modelos independentes sempre obtiveram piores ajustes tanto quando modelados os dados com respostas binárias, quanto com respostas multinomiais. Quando consideramos  $K$  posições dependentes, o número de parâmetros de dependência de cada modelo aumenta diferentemente, como mostra a Tabela 7.1. O total de interceptos será sempre  $3K$  para todos os modelos, assim como o total de parâmetros para as covariáveis será sempre 3, nesse caso.

Tabela 7.1: Número de Parâmetros dos Modelos Multinomiais Multivariados para  $K$  Posições Dependentes

Modelo	Total de Parâmetros de Dependência
Independente	0
Ig. Preditivo	0 para $K = 1$ ; 3 para $K > 1$
Markov 1ª Ordem	$3(K - 1)$
Aditivo	$3 \times K(K - 1)/2$
Locação	$K(K - 1)/2$
Transição	0 para $K = 1$ ; 9 para $K > 1$
Semi-Locação e Transição	$9(K - 1)$
Locação e Transição	$9 \times K(K - 1)/2$

A Tabela 7.2 apresenta o total de parâmetros de dependência para todos os modelos, conforme aumenta-se o número de posições dependentes, lembrando que cada seqüência do gene NADH4 possui 459 códons efetivos, há portanto 1377 posições a serem modeladas em cada seqüência.

Pela Tabela 7.2, vê-se que o modelo aditivo terá 2842128 parâmetros de dependência quando toda a seqüência for modelada, 229,5 vezes mais parâmetros do que o modelo de semi-locação e transição, que terá 12384. Se levado em consideração que é muito provável que em modelos para toda a seqüência, os valores das log-verossimilhanças dos modelos sejam muito próximos, a penalização do número de parâmetros para o cálculo

Tabela 7.2: Total de Parâmetros de Dependência dos Modelos Multinomiais Multivariados Conforme o Número  $K$  de Posições Dependentes Aumenta

Modelo	Total de Posições Dependentes										
	1	2	3	4	5	6	7	...	100	...	1377
Independente	0	0	0	0	0	0	0	...	0	...	0
Ig. Preditivo	0	3	3	3	3	3	3	...	3	...	3
Markov 1 <sup>a</sup> Ordem	0	3	6	9	12	15	18	...	297	...	4128
Aditivo	0	3	9	18	30	45	63	...	14850	...	2842128
Locação	0	1	3	6	10	15	21	...	4950	...	947376
Transição	0	9	9	9	9	9	9	...	9	...	9
Semi-Loc. e Trans.	0	9	18	27	36	45	54	...	891	...	12384
Locação e Transição	0	9	27	54	90	135	189	...	44550	...	8526384

de medidas como o AIC e BIC no modelo aditivo será muito maior do que no modelo de semi-locação e transio.

É interessante discutir também o porquê de não se fazer o uso da estrutura markoviana de primeira ordem, uma vez que este modelo também possui menor AIC, BIC e SQE que o modelo de semi-locação e transição. Para 1377 posições dependentes o modelo markoviano de primeira ordem possui 4128 parâmetros de dependência. A diferença desse modelo é exatamente o fato de ser uma estrutura markoviana de primeira ordem, ou seja, não inclui a informação referente a toda seqüência anterior à posição que estiver sendo modelada, como fazem os modelos aditivo e de semi-locação e transição, mas apenas a posição imediatamente anterior.

Assim, para a modelagem da probabilidade dos códons, o modelo aditivo possui melhor ajuste. No entanto, ao aumentar a estrutura de dependência para além dos códons, por exemplo para todo o gene, conforme é mostrado na Tabela 7.2, o número de parâmetros do modelo aditivo cresce muito mais que no modelo de semi-locação e transição, e devido ao fato do AIC e BIC do modelo baseado na representação de

---

Bahadur não serem proporcionalmente muito maiores do que os do modelo regressivo, talvez o modelo de semi-locação e transição seja mais adequado e apresente melhor ajuste, uma vez que seu número de parâmetros será muito menor.

É importante ressaltar que neste trabalho os códons foram considerados independentes entre si, assim como em [Bonney et al. \(1994\)](#), o que provavelmente não seja uma verdade. A modelagem dos códons sob a suposição de independência facilitou o desenvolvimento e aplicação de todos os modelos aqui propostos e estudados. No entanto, é fundamental que em trabalhos futuros, esses modelos sejam expandidos não mais para a modelagem de frequências de códons em um gene, mas sim para a frequência das bases nitrogenadas em cada posição do gene. É justamente neste caso que o modelo baseado na representação de Bahadur de semi-locação e transição se destaca. Quando tratando apenas de códons, o modelo aditivo possui um menor número de parâmetros, porém quando expandidos os modelos para toda a seqüência de um gene, o número de parâmetros do modelo de semi-locação e transição é deveras inferior, como mostrado na Tabela [7.2](#).

Outra questão a ser explorada em trabalhos futuros é a questão das técnicas computacionais para estimação dos parâmetros dos modelos. Neste trabalho, por serem estudados apenas os códons, considerados independentes entre si, o banco de dados com 30 seqüências do gene NADH4 possui um total de 13770 códons, sendo portanto, o tamanho da amostra maior do que o número de parâmetros de todos os modelos propostos. Quando expandidos para toda a seqüência, os modelos apresentarão um número muito grande de parâmetros, o que possivelmente causará dificuldades computacionais. É necessário também em modelos envolvendo a estrutura de dependência de todo um gene, a busca de novas covariáveis, uma vez que as aqui apresentadas são intrinsicamente relacionadas aos códons e não ao gene como um todo.

# Referências Bibliográficas

- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle *Second International Symposium of Information Theory* B.N. Petrov and F. Csaki (eds) Budapest, Akademia Kiado 267–281
- Alberts, B.; Johnson, A.; Lewis J.; Raff, M.; Roberts, K. and Walter, P. (2002). Molecular Biology of the Cell. *Garland Science, Second Edition*
- Anderson, S.; Bankier, A. and Barrell, B. (1981). Sequence and Organization of the Human Mitochondrial Genome. *Nature* **290** 457–465
- Andrews, R.; Kubacka, I.; Chinnery, P.; Lightowers, R.; Turnbull, D. and Howell, N. (1999). Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA. *Nature Genet.* **23** 147
- Agresti, A. (2002). Categorical Data Analysis. *Wiley Series in Probability and Mathematical Statistics, Second Edition. Applied Probability and Statistics, Hardcover*
- Bahadur, R.R. (1961). A Representation of the Joint Distribution of Responses to n Dichotomous Items. *In studies in Item Analysis and Prediction* 158–176
- Bonney, G.E. (1986). Regressive Logistic Models for Familial Disease and Other Binary Traits. *Biometrics* **42** 611–625
- Bonney, G.E. (1987). Logistic Regression for Dependent Binary Observations. *Biometrics* **43** 951–973



- Bonney, G.E.; Dunston, G. and Wilson, J. (1989). The Use of Regressive Logistic Models for Ordered and Unordered Polytomous Traits: Application to Affective Disorders. *Genetics Epidemiology* **6** 211–215
- Bonney, G.E.; Amfoh, K. and Shaw, R. (1994). The Use of Logistic Models for the Analysis of Codon Frequencies of DNA Sequences in Terms of Explanatory Variables. *Biometrics* **50** 1054–1063
- Casella, R. and Berger, G.L. (2002). Statistical Inference. *Duxbury Advanced Series*
- Cox, D.R. (1972). The Analysis of Multivariate Binary Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **Vol. 21, No. 2** 113–120
- Fitzmaurice, G.M.; Laird, N.M. and Rotnitzky, A.G. (1993). Regression Models for Discrete Longitudinal Responses. *Statistical Science* **Vol. 8, No. 3** 284–299
- Fitzmaurice, G.M. (1995). A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data. *Biometrics* **Vol. 51, No. 1** 309–317
- Grandin, L.C. (2006). Aplicações de Modelos Logísticos Regressivos em Biologia Molecular. *Dissertação de Mestrado Apresentada junto ao Departamento de Estatística, da Universidade Estadual de Campinas*
- Grantham, R. (1974). Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185** 862–864
- Hoffman, K.M. and Kunze, R. (1971). Linear Algebra. *Prentice Hall, Second Edition*
- Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression. *Wiley Series in Probability and Statistics*
- Lesaffre, E. and Albert, A. (1989). Multiple-group Logistic Regression Diagnostics. *Applied Statistics* **Vol. 38, No. 3** 425–440

- National Center for Biotechnology Information* - NCBI (2010), <http://www.ncbi.nlm.nih.gov/>
- Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society Series A (General)* **Vol. 135, No. 3** 370–384
- Nyangoma, S.O; Fung, W-K. and Jansen, R.C. (2006). Identifying Influential Multinomial Observations by Perturbation. *Computational Statistics & Data Analysis* 50 2799–2821
- Parzen, M.; Ghosh, S.; Lipsitz, S.R.; Sinha, D.; Fitzmaurice, G.M.; Ibrahim, J.G. and Mallick, B.K. (2009). A Generalized Linear Mixed Model for Longitudinal Binary Data with a Marginal Logit Link Function. *Conditionally accepted for publication in the Journal of the Royal Statistical Society, Series B*
- Paula, G. (2004). Modelos de Regressão com Apoio Computacional. *Notas de aula, Universidade de São Paulo* <http://www.ime.usp.br/~giapaula>
- Pinheiro, H.P.; Seillier-Moiseiwitsch, F. and Sen, P.K. (1999). Modeling the Mutation Process in the HIV Genome. *Research Report #13/1999. University of Campinas, Brazil*
- Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics* 9, No. 4 705–724
- Raftery, A.E. (1986). Choosing Models for Cross-Classification. *American Sociological Review* **Vol. 51, No. 1** 145–146
- Reilly, C. (2009). Statistics in Human Genetic and Molecular Biology. *Chapman & Hall/CRC Texts in Statistical Science Series*
- The R Project for Statistical Computing* (2010), <http://www.r-project.org/>
- Seber, G.A.F. and Nyangoma, S.O. (2000). Residuals for Multinomial Models. *Biometrika* **Vol. 87, No. 1** 183–191

- 
- Stefanescu, C. and Turnbull, B.W. (2003). Likelihood Inference for Exchangeable Binary Data with Varying Cluster Sizes. *Biometrics* **Vol. 59, No. 1** 18–24
- Watson, J.D. and Crick, F (1953). Molecular Structure of Nucleic Acids. *Nature* **171** 737–738
- Zhao, L.P. and Prentice, R.L (1990). Correlated Binary Regression Using a Quadratic Exponential Model. *Biometrika* **Vol. 77, No. 3** 642–648

# Apêndice 1

## Código para o Software R

O código apresentado a seguir é referente à log-verossimilhança dos modelos multinomiais regressivos independente e aditivo, e do segundo passo do modelo baseado na representação de Bahadur de semi-locação & transição. Os demais códigos e o banco de dados utilizados nas implementações e análises dos modelos, por serem muito extensos, não estão incluídos nesta dissertação, devido ao espaço que ocupariam. No entanto, podem ser solicitados às autoras através dos emails [bia.cdc@gmail.com](mailto:bia.cdc@gmail.com) ou [hildete@ime.unicamp.br](mailto:hildete@ime.unicamp.br).

```
## REGRESSIVE - INDEPENDENT ##
loglike1 <- function(u)
{
a11 <- u[1]
a12 <- u[2]
a13 <- u[3]
a21 <- u[4]
a22 <- u[5]
a23 <- u[6]
a31 <- u[7]
a32 <- u[8]
a33 <- u[9]
b1 <- u[10]
b2 <- u[11]
b3 <- u[12]

t11 <- a11 + b1*aarisk + b2*avdist + b3*tscore
t12 <- a12 + b1*aarisk + b2*avdist + b3*tscore
t13 <- a13 + b1*aarisk + b2*avdist + b3*tscore
t21 <- a21 + b1*aarisk + b2*avdist + b3*tscore
t22 <- a22 + b1*aarisk + b2*avdist + b3*tscore
t23 <- a23 + b1*aarisk + b2*avdist + b3*tscore
t31 <- a31 + b1*aarisk + b2*avdist + b3*tscore
t32 <- a32 + b1*aarisk + b2*avdist + b3*tscore
t33 <- a33 + b1*aarisk + b2*avdist + b3*tscore
p1 <- exp(z11*t11 + z12*t12 + z13*t13)/(1 + exp(t11) + exp(t12) + exp(t13))
}
```

```

p2 <- exp(z21*t21 + z22*t22 + z23*t23)/(1 + exp(t21) + exp(t22) + exp(t23))
p3 <- exp(z31*t31 + z32*t32 + z33*t33)/(1 + exp(t31) + exp(t32) + exp(t33))
pc <- p1*p2*p3
l <- n*(log(pc) - log(sum(pc)))
loglike <- sum(l)
return(-loglike) # R's optim() routines minimizes objective functions
}
grad1 <- function(u)
{
a11 <- u[1]
a12 <- u[2]
a13 <- u[3]
a21 <- u[4]
a22 <- u[5]
a23 <- u[6]
a31 <- u[7]
a32 <- u[8]
a33 <- u[9]
b1 <- u[10]
b2 <- u[11]
b3 <- u[12]

t11 <- a11 + b1*aarisk + b2*avdist + b3*tscore
t12 <- a12 + b1*aarisk + b2*avdist + b3*tscore
t13 <- a13 + b1*aarisk + b2*avdist + b3*tscore
t21 <- a21 + b1*aarisk + b2*avdist + b3*tscore
t22 <- a22 + b1*aarisk + b2*avdist + b3*tscore
t23 <- a23 + b1*aarisk + b2*avdist + b3*tscore
t31 <- a31 + b1*aarisk + b2*avdist + b3*tscore
t32 <- a32 + b1*aarisk + b2*avdist + b3*tscore
t33 <- a33 + b1*aarisk + b2*avdist + b3*tscore

p1 <- exp(z11*t11 + z12*t12 + z13*t13)/(1 + exp(t11) + exp(t12) + exp(t13))
p2 <- exp(z21*t21 + z22*t22 + z23*t23)/(1 + exp(t21) + exp(t22) + exp(t23))
p3 <- exp(z31*t31 + z32*t32 + z33*t33)/(1 + exp(t31) + exp(t32) + exp(t33))
pc <- p1*p2*p3

dlp_a11 <- z11 - exp(t11)/(1 + exp(t11) + exp(t12) + exp(t13))
dlp_a12 <- z12 - exp(t12)/(1 + exp(t11) + exp(t12) + exp(t13))
dlp_a13 <- z13 - exp(t13)/(1 + exp(t11) + exp(t12) + exp(t13))
dlp_a21 <- z21 - exp(t21)/(1 + exp(t21) + exp(t22) + exp(t23))
dlp_a22 <- z22 - exp(t22)/(1 + exp(t21) + exp(t22) + exp(t23))
dlp_a23 <- z23 - exp(t23)/(1 + exp(t21) + exp(t22) + exp(t23))
dlp_a31 <- z31 - exp(t31)/(1 + exp(t31) + exp(t32) + exp(t33))
dlp_a32 <- z32 - exp(t32)/(1 + exp(t31) + exp(t32) + exp(t33))
dlp_a33 <- z33 - exp(t33)/(1 + exp(t31) + exp(t32) + exp(t33))

dlp_b1 <- (dlp_a11 + dlp_a12 + dlp_a13 + dlp_a21 + dlp_a22 + dlp_a23 + dlp_a31 + dlp_a32 + dlp_a33)*aarisk
dlp_b2 <- (dlp_a11 + dlp_a12 + dlp_a13 + dlp_a21 + dlp_a22 + dlp_a23 + dlp_a31 + dlp_a32 + dlp_a33)*avdist
dlp_b3 <- (dlp_a11 + dlp_a12 + dlp_a13 + dlp_a21 + dlp_a22 + dlp_a23 + dlp_a31 + dlp_a32 + dlp_a33)*tscore

dlsum_a11 <- (1/sum(pc))*sum(pc*dlp_a11)
dlsum_a12 <- (1/sum(pc))*sum(pc*dlp_a12)
dlsum_a13 <- (1/sum(pc))*sum(pc*dlp_a13)
dlsum_a21 <- (1/sum(pc))*sum(pc*dlp_a21)
dlsum_a22 <- (1/sum(pc))*sum(pc*dlp_a22)
dlsum_a23 <- (1/sum(pc))*sum(pc*dlp_a23)
dlsum_a31 <- (1/sum(pc))*sum(pc*dlp_a31)
dlsum_a32 <- (1/sum(pc))*sum(pc*dlp_a32)
dlsum_a33 <- (1/sum(pc))*sum(pc*dlp_a33)
dlsum_b1 <- (1/sum(pc))*sum(pc*dlp_b1)
dlsum_b2 <- (1/sum(pc))*sum(pc*dlp_b2)
dlsum_b3 <- (1/sum(pc))*sum(pc*dlp_b3)
da11 <- n*(dlp_a11 - dlsum_a11)

```

```

da12 <- n*(dlp_a12 - dlsum_a12)
da13 <- n*(dlp_a13 - dlsum_a13)
da21 <- n*(dlp_a21 - dlsum_a21)
da22 <- n*(dlp_a22 - dlsum_a22)
da23 <- n*(dlp_a23 - dlsum_a23)
da31 <- n*(dlp_a31 - dlsum_a31)
da32 <- n*(dlp_a32 - dlsum_a32)
da33 <- n*(dlp_a33 - dlsum_a33)
db1 <- n*(dlp_b1 - dlsum_b1)
db2 <- n*(dlp_b2 - dlsum_b2)
db3 <- n*(dlp_b3 - dlsum_b3)
grad <- c(sum(da11),sum(da12),sum(da13),sum(da21),sum(da22),sum(da23),sum(da31),sum(da32),sum(da33),sum(db1),sum(db2),sum(db3))
return(-grad)
}
## REGRESSIVE - ADDITIVE ##
loglike4 <- function(u)
{
a11 <- u[1]
a12 <- u[2]
a13 <- u[3]
a21 <- u[4]
a22 <- u[5]
a23 <- u[6]
a31 <- u[7]
a32 <- u[8]
a33 <- u[9]
g11 <- u[10]
g12 <- u[11]
g13 <- u[12]
g21 <- u[13]
g22 <- u[14]
g23 <- u[15]
g31 <- u[16]
g32 <- u[17]
g33 <- u[18]
b1 <- u[19]
b2 <- u[20]
b3 <- u[21]
t11 <- a11 + b1*aarisk + b2*avdist + b3*tscore
t12 <- a12 + b1*aarisk + b2*avdist + b3*tscore
t13 <- a13 + b1*aarisk + b2*avdist + b3*tscore
t21 <- a21 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t22 <- a22 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t23 <- a23 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t31 <- a31 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
t32 <- a32 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
t33 <- a33 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
p1 <- exp(z11*t11 + z12*t12 + z13*t13)/(1 + exp(t11) + exp(t12) + exp(t13))
p2 <- exp(z21*t21 + z22*t22 + z23*t23)/(1 + exp(t21) + exp(t22) + exp(t23))
p3 <- exp(z31*t31 + z32*t32 + z33*t33)/(1 + exp(t31) + exp(t32) + exp(t33))
pc <- p1*p2*p3
l <- n*(log(pc) - log(sum(pc)))
loglike <- sum(l)
return(-loglike) # R's optim() routines minimizes objective functions
}
grad4 <- function(u)
{
a11 <- u[1]
a12 <- u[2]

```

```

a13 <- u[3]
a21 <- u[4]
a22 <- u[5]
a23 <- u[6]
a31 <- u[7]
a32 <- u[8]
a33 <- u[9]
g11 <- u[10]
g12 <- u[11]
g13 <- u[12]
g21 <- u[13]
g22 <- u[14]
g23 <- u[15]
g31 <- u[16]
g32 <- u[17]
g33 <- u[18]
b1 <- u[19]
b2 <- u[20]
b3 <- u[21]

t11 <- a11 + b1*aarisk + b2*avdist + b3*tscore
t12 <- a12 + b1*aarisk + b2*avdist + b3*tscore
t13 <- a13 + b1*aarisk + b2*avdist + b3*tscore
t21 <- a21 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t22 <- a22 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t23 <- a23 + g11*z11 + g12*z12 + g13*z13 + b1*aarisk + b2*avdist + b3*tscore
t31 <- a31 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
t32 <- a32 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
t33 <- a33 + g21*z11 + g22*z12 + g23*z13 + g31*z21 + g32*z22 + g33*z23 + b1*aarisk + b2*avdist + b3*tscore
p1 <- exp(z11*t11 + z12*t12 + z13*t13)/(1 + exp(t11) + exp(t12) + exp(t13))
p2 <- exp(z21*t21 + z22*t22 + z23*t23)/(1 + exp(t21) + exp(t22) + exp(t23))
p3 <- exp(z31*t31 + z32*t32 + z33*t33)/(1 + exp(t31) + exp(t32) + exp(t33))
pc <- p1*p2*p3

d1p_a11 <- z11 - exp(t11)/(1 + exp(t11) + exp(t12) + exp(t13))
d1p_a12 <- z12 - exp(t12)/(1 + exp(t11) + exp(t12) + exp(t13))
d1p_a13 <- z13 - exp(t13)/(1 + exp(t11) + exp(t12) + exp(t13))
d1p_a21 <- z21 - exp(t21)/(1 + exp(t21) + exp(t22) + exp(t23))
d1p_a22 <- z22 - exp(t22)/(1 + exp(t21) + exp(t22) + exp(t23))
d1p_a23 <- z23 - exp(t23)/(1 + exp(t21) + exp(t22) + exp(t23))
d1p_a31 <- z31 - exp(t31)/(1 + exp(t31) + exp(t32) + exp(t33))
d1p_a32 <- z32 - exp(t32)/(1 + exp(t31) + exp(t32) + exp(t33))
d1p_a33 <- z33 - exp(t33)/(1 + exp(t31) + exp(t32) + exp(t33))

d1p_g11 <- (d1p_a21 + d1p_a22 + d1p_a23)*z11
d1p_g12 <- (d1p_a21 + d1p_a22 + d1p_a23)*z12
d1p_g13 <- (d1p_a21 + d1p_a22 + d1p_a23)*z13
d1p_g21 <- (d1p_a31 + d1p_a32 + d1p_a33)*z11
d1p_g22 <- (d1p_a31 + d1p_a32 + d1p_a33)*z12
d1p_g23 <- (d1p_a31 + d1p_a32 + d1p_a33)*z13
d1p_g31 <- (d1p_a31 + d1p_a32 + d1p_a33)*z21
d1p_g32 <- (d1p_a31 + d1p_a32 + d1p_a33)*z22
d1p_g33 <- (d1p_a31 + d1p_a32 + d1p_a33)*z23

d1p_b1 <- (d1p_a11 + d1p_a12 + d1p_a13 + d1p_a21 + d1p_a22 + d1p_a23 + d1p_a31 + d1p_a32 + d1p_a33)*aarisk
d1p_b2 <- (d1p_a11 + d1p_a12 + d1p_a13 + d1p_a21 + d1p_a22 + d1p_a23 + d1p_a31 + d1p_a32 + d1p_a33)*avdist
d1p_b3 <- (d1p_a11 + d1p_a12 + d1p_a13 + d1p_a21 + d1p_a22 + d1p_a23 + d1p_a31 + d1p_a32 + d1p_a33)*tscore
d1sum_a11 <- (1/sum(pc))*sum(pc*d1p_a11)
d1sum_a12 <- (1/sum(pc))*sum(pc*d1p_a12)
d1sum_a13 <- (1/sum(pc))*sum(pc*d1p_a13)
d1sum_a21 <- (1/sum(pc))*sum(pc*d1p_a21)
d1sum_a22 <- (1/sum(pc))*sum(pc*d1p_a22)
d1sum_a23 <- (1/sum(pc))*sum(pc*d1p_a23)

```

```

dlsun_a31 <- (1/sum(pc))*sum(pc*d1p_a31)
dlsun_a32 <- (1/sum(pc))*sum(pc*d1p_a32)
dlsun_a33 <- (1/sum(pc))*sum(pc*d1p_a33)
dlsun_g11 <- (1/sum(pc))*sum(pc*d1p_g11)
dlsun_g12 <- (1/sum(pc))*sum(pc*d1p_g12)
dlsun_g13 <- (1/sum(pc))*sum(pc*d1p_g13)
dlsun_g21 <- (1/sum(pc))*sum(pc*d1p_g21)
dlsun_g22 <- (1/sum(pc))*sum(pc*d1p_g22)
dlsun_g23 <- (1/sum(pc))*sum(pc*d1p_g23)
dlsun_g31 <- (1/sum(pc))*sum(pc*d1p_g31)
dlsun_g32 <- (1/sum(pc))*sum(pc*d1p_g32)
dlsun_g33 <- (1/sum(pc))*sum(pc*d1p_g33)
dlsun_b1 <- (1/sum(pc))*sum(pc*d1p_b1)
dlsun_b2 <- (1/sum(pc))*sum(pc*d1p_b2)
dlsun_b3 <- (1/sum(pc))*sum(pc*d1p_b3)
da11 <- n*(d1p_a11 - dlsun_a11)
da12 <- n*(d1p_a12 - dlsun_a12)
da13 <- n*(d1p_a13 - dlsun_a13)
da21 <- n*(d1p_a21 - dlsun_a21)
da22 <- n*(d1p_a22 - dlsun_a22)
da23 <- n*(d1p_a23 - dlsun_a23)
da31 <- n*(d1p_a31 - dlsun_a31)
da32 <- n*(d1p_a32 - dlsun_a32)
da33 <- n*(d1p_a33 - dlsun_a33)
dg11 <- n*(d1p_g11 - dlsun_g11)
dg12 <- n*(d1p_g12 - dlsun_g12)
dg13 <- n*(d1p_g13 - dlsun_g13)
dg21 <- n*(d1p_g21 - dlsun_g21)
dg22 <- n*(d1p_g22 - dlsun_g22)
dg23 <- n*(d1p_g23 - dlsun_g23)
dg31 <- n*(d1p_g31 - dlsun_g31)
dg32 <- n*(d1p_g32 - dlsun_g32)
dg33 <- n*(d1p_g33 - dlsun_g33)
db1 <- n*(d1p_b1 - dlsun_b1)
db2 <- n*(d1p_b2 - dlsun_b2)
db3 <- n*(d1p_b3 - dlsun_b3)
grad <- c(sum(da11),sum(da12),sum(da13),sum(da21),sum(da22),sum(da23),sum(da31),sum(da32),sum(da33),sum(dg11),sum(dg12),
sum(dg13),sum(dg21),sum(dg22),sum(dg23),sum(dg31),sum(dg32),sum(dg33),sum(db1),sum(db2),sum(db3))
return(-grad)
}
## BAHADUR - SEMI LOCATION & TRANSITION ##
loglike7 <- function(u)
{
r1.11 <- u[1]
r1.12 <- u[2]
r1.13 <- u[3]
r1.21 <- u[4]
r1.22 <- u[5]
r1.23 <- u[6]
r1.31 <- u[7]
r1.32 <- u[8]
r1.33 <- u[9]
r2.11 <- u[10]
r2.12 <- u[11]
r2.13 <- u[12]
r2.21 <- u[13]
r2.22 <- u[14]
r2.23 <- u[15]
r2.31 <- u[16]

```



```

r2.32 <- u[17]
r2.33 <- u[18]
f <- 1 + r1.11*(u11*u21+u21*u31) + r1.12*(u11*u22+u21*u32) + r1.13*(u11*u23+u21*u33) + r1.21*(u12*u21+u22*u31) +
  r1.22*(u12*u22+u22*u32) + r1.23*(u12*u23+u22*u33) + r1.31*(u13*u21+u23*u31) + r1.32*(u13*u22+u23*u32) + r1.33*(u13*u23+u23*u33) +
  r2.11*(u11*u31) + r2.12*(u11*u32) + r2.13*(u11*u33) + r2.21*(u12*u31) + r2.22*(u12*u32) + r2.23*(u12*u33) + r2.31*(u13*u31) +
  r2.32*(u13*u32) + r2.33*(u13*u33)
l <- n*(log(f) - log(sum(p_ind*f)))
loglike <- sum(l)
return(-loglike) # R's constrOptim() routines minimizes objective functions
}
grad7 <- function(u)
{
r1.11 <- u[1]
r1.12 <- u[2]
r1.13 <- u[3]
r1.21 <- u[4]
r1.22 <- u[5]
r1.23 <- u[6]
r1.31 <- u[7]
r1.32 <- u[8]
r1.33 <- u[9]
r2.11 <- u[10]
r2.12 <- u[11]
r2.13 <- u[12]
r2.21 <- u[13]
r2.22 <- u[14]
r2.23 <- u[15]
r2.31 <- u[16]
r2.32 <- u[17]
r2.33 <- u[18]
f <- 1 + r1.11*(u11*u21+u21*u31) + r1.12*(u11*u22+u21*u32) + r1.13*(u11*u23+u21*u33) + r1.21*(u12*u21+u22*u31) +
  r1.22*(u12*u22+u22*u32) + r1.23*(u12*u23+u22*u33) + r1.31*(u13*u21+u23*u31) + r1.32*(u13*u22+u23*u32) + r1.33*(u13*u23+u23*u33) +
  r2.11*(u11*u31) + r2.12*(u11*u32) + r2.13*(u11*u33) + r2.21*(u12*u31) + r2.22*(u12*u32) + r2.23*(u12*u33) + r2.31*(u13*u31) +
  r2.32*(u13*u32) + r2.33*(u13*u33)
dr1.11 <- n*((u11*u21+u21*u31)/f) - (sum(p_ind*(u11*u21+u21*u31))/sum(p_ind*f))
dr1.12 <- n*((u11*u22+u21*u32)/f) - (sum(p_ind*(u11*u22+u21*u32))/sum(p_ind*f))
dr1.13 <- n*((u11*u23+u21*u33)/f) - (sum(p_ind*(u11*u23+u21*u33))/sum(p_ind*f))
dr1.21 <- n*((u12*u21+u22*u31)/f) - (sum(p_ind*(u12*u21+u22*u31))/sum(p_ind*f))
dr1.22 <- n*((u12*u22+u22*u32)/f) - (sum(p_ind*(u12*u22+u22*u32))/sum(p_ind*f))
dr1.23 <- n*((u12*u23+u22*u33)/f) - (sum(p_ind*(u12*u23+u22*u33))/sum(p_ind*f))
dr1.31 <- n*((u13*u21+u23*u31)/f) - (sum(p_ind*(u13*u21+u23*u31))/sum(p_ind*f))
dr1.32 <- n*((u13*u22+u23*u32)/f) - (sum(p_ind*(u13*u22+u23*u32))/sum(p_ind*f))
dr1.33 <- n*((u13*u23+u23*u33)/f) - (sum(p_ind*(u13*u23+u23*u33))/sum(p_ind*f))
dr2.11 <- n*((u11*u31)/f) - (sum(p_ind*(u11*u31))/sum(p_ind*f))
dr2.12 <- n*((u11*u32)/f) - (sum(p_ind*(u11*u32))/sum(p_ind*f))
dr2.13 <- n*((u11*u33)/f) - (sum(p_ind*(u11*u33))/sum(p_ind*f))
dr2.21 <- n*((u12*u31)/f) - (sum(p_ind*(u12*u31))/sum(p_ind*f))
dr2.22 <- n*((u12*u32)/f) - (sum(p_ind*(u12*u32))/sum(p_ind*f))
dr2.23 <- n*((u12*u33)/f) - (sum(p_ind*(u12*u33))/sum(p_ind*f))
dr2.31 <- n*((u13*u31)/f) - (sum(p_ind*(u13*u31))/sum(p_ind*f))
dr2.32 <- n*((u13*u32)/f) - (sum(p_ind*(u13*u32))/sum(p_ind*f))
dr2.33 <- n*((u13*u33)/f) - (sum(p_ind*(u13*u33))/sum(p_ind*f))
grad <- c(sum(dr1.11),sum(dr1.12),sum(dr1.13),sum(dr1.21),sum(dr1.22),sum(dr1.23),sum(dr1.31),sum(dr1.32),sum(dr1.33),sum(dr2.11),
  sum(dr2.12),sum(dr2.13),sum(dr2.21),sum(dr2.22),sum(dr2.23),sum(dr2.31),sum(dr2.32),sum(dr2.33))
return(-grad)
}

```

# Apêndice 2

## Probabilidades Estimadas pelos Modelos

As tabelas a seguir apresentam as probabilidades estimadas para todos os códons efetivos nas seqüências do gene NADH4, para todos os modelos propostos, binomiais e multinomiais logísticos regressivos e baseados na representação de Bahadur.

Tabela 7.3: Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Logito

Códon	Obs.	Modelo Independente	Modelo Ig. Preditivo	Estrutura Markoviana	Modelo Aditivo	Modelo Bahadur
TTT	0.0196	0.0308	0.0292	0.0303	0.0293	0.0293
CTT	0.0219	0.0350	0.0332	0.0357	0.0340	0.0335
ATT	0.0349	0.0250	0.0265	0.0237	0.0248	0.0254
GTT	0.0000	0.0236	0.0251	0.0223	0.0234	0.0240
TCT	0.0109	0.0281	0.0266	0.0267	0.0263	0.0266
CCT	0.0055	0.0225	0.0198	0.0215	0.0204	0.0210
ACT	0.0174	0.0192	0.0198	0.0175	0.0185	0.0194
GCT	0.0131	0.0266	0.0277	0.0253	0.0262	0.0271
TAT	0.0045	0.0080	0.0074	0.0077	0.0073	0.0090
CAT	0.0022	0.0088	0.0094	0.0104	0.0102	0.0098
AAT	0.0027	0.0148	0.0157	0.0167	0.0170	0.0147
GAT	0.0000	0.0095	0.0086	0.0095	0.0093	0.0093
TGT	0.0022	0.0084	0.0077	0.0082	0.0075	0.0090
CGT	0.0000	0.0096	0.0107	0.0113	0.0113	0.0107
AGT	0.0044	0.0119	0.0137	0.0146	0.0158	0.0118
GGT	0.0022	0.0097	0.0090	0.0096	0.0095	0.0095
TTC	0.0240	0.0308	0.0292	0.0303	0.0293	0.0293
CTC	0.0674	0.0350	0.0332	0.0357	0.0340	0.0335
ATC	0.0501	0.0250	0.0265	0.0237	0.0248	0.0254
GTC	0.0087	0.0236	0.0251	0.0223	0.0234	0.0240
TCC	0.0370	0.0281	0.0266	0.0267	0.0263	0.0266
CCC	0.0316	0.0225	0.0198	0.0215	0.0204	0.0210
ACC	0.0371	0.0192	0.0198	0.0175	0.0185	0.0194
GCC	0.0261	0.0266	0.0277	0.0253	0.0262	0.0271
TAC	0.0238	0.0080	0.0074	0.0077	0.0073	0.0090
CAC	0.0266	0.0088	0.0094	0.0104	0.0102	0.0098
AAC	0.0474	0.0148	0.0157	0.0167	0.0170	0.0147
GAC	0.0065	0.0095	0.0086	0.0095	0.0093	0.0093
TGC	0.0044	0.0084	0.0077	0.0082	0.0075	0.0090
CGC	0.0105	0.0096	0.0107	0.0113	0.0113	0.0107
AGC	0.0174	0.0119	0.0137	0.0146	0.0158	0.0118
GGC	0.0196	0.0097	0.0090	0.0096	0.0095	0.0095
TTA	0.0173	0.0180	0.0189	0.0190	0.0200	0.0196
CTA	0.0918	0.0179	0.0190	0.0190	0.0201	0.0196
ATA	0.0523	0.0184	0.0186	0.0201	0.0193	0.0177
GTA	0.0174	0.0187	0.0180	0.0194	0.0182	0.0181
TCA	0.0218	0.0178	0.0185	0.0185	0.0195	0.0195
CCA	0.0131	0.0168	0.0172	0.0176	0.0184	0.0185
ACA	0.0480	0.0238	0.0245	0.0258	0.0248	0.0234
GCA	0.0174	0.0134	0.0119	0.0133	0.0123	0.0128
CAA	0.0196	0.0120	0.0131	0.0102	0.0108	0.0106
AAA	0.0219	0.0132	0.0130	0.0131	0.0129	0.0135
GAA	0.0195	0.0115	0.0098	0.0098	0.0091	0.0119
TGA	0.0261	0.0092	0.0103	0.0084	0.0088	0.0078
CGA	0.0087	0.0123	0.0133	0.0106	0.0111	0.0110
GGA	0.0107	0.0050	0.0036	0.0037	0.0034	0.0054
TTG	0.0022	0.0180	0.0189	0.0190	0.0200	0.0196
CTG	0.0086	0.0179	0.0190	0.0190	0.0201	0.0196
ATG	0.0065	0.0184	0.0186	0.0201	0.0193	0.0177
GTG	0.0022	0.0187	0.0180	0.0194	0.0182	0.0181
TCG	0.0022	0.0178	0.0185	0.0185	0.0195	0.0195
CCG	0.0000	0.0168	0.0172	0.0176	0.0184	0.0185
ACG	0.0021	0.0238	0.0245	0.0258	0.0248	0.0234
GCG	0.0000	0.0134	0.0119	0.0133	0.0123	0.0128
CAG	0.0022	0.0131	0.0138	0.0113	0.0117	0.0119
AAG	0.0022	0.0132	0.0130	0.0131	0.0129	0.0135
GAG	0.0000	0.0115	0.0098	0.0098	0.0091	0.0119
TGG	0.0022	0.0092	0.0103	0.0084	0.0088	0.0078
CGG	0.0000	0.0123	0.0133	0.0106	0.0111	0.0110
GGG	0.0046	0.0050	0.0036	0.0037	0.0034	0.0054

Tabela 7.4: Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Probita

Códon	Obs.	Modelo Independente	Modelo Ig. Preditivo	Estrutura Markoviana	Modelo Aditivo	Modelo Bahadur
TTT	0.0196	0.0258	0.0242	0.0247	0.0248	0.0255
CTT	0.0219	0.0348	0.0328	0.0351	0.0355	0.0344
ATT	0.0349	0.0245	0.0256	0.0239	0.0237	0.0235
GTT	0.0000	0.0244	0.0255	0.0238	0.0236	0.0234
TCT	0.0109	0.0200	0.0187	0.0182	0.0181	0.0197
CCT	0.0055	0.0276	0.0266	0.0277	0.0279	0.0273
ACT	0.0174	0.0243	0.0255	0.0238	0.0236	0.0232
GCT	0.0131	0.0252	0.0259	0.0243	0.0241	0.0242
TAT	0.0045	0.0072	0.0066	0.0071	0.0071	0.0078
CAT	0.0022	0.0102	0.0112	0.0121	0.0122	0.0110
AAT	0.0027	0.0131	0.0126	0.0133	0.0133	0.0135
GAT	0.0000	0.0122	0.0119	0.0125	0.0125	0.0126
TGT	0.0022	0.0079	0.0079	0.0076	0.0075	0.0086
CGT	0.0000	0.0093	0.0104	0.0111	0.0111	0.0101
AGT	0.0044	0.0119	0.0137	0.0136	0.0134	0.0123
GGT	0.0022	0.0108	0.0103	0.0106	0.0106	0.0112
TTC	0.0240	0.0258	0.0242	0.0247	0.0248	0.0255
CTC	0.0674	0.0348	0.0328	0.0351	0.0355	0.0344
ATC	0.0501	0.0245	0.0256	0.0239	0.0237	0.0235
GTC	0.0087	0.0244	0.0255	0.0238	0.0236	0.0234
TCC	0.0370	0.0200	0.0187	0.0182	0.0181	0.0197
CCC	0.0316	0.0276	0.0266	0.0277	0.0279	0.0273
ACC	0.0371	0.0243	0.0255	0.0238	0.0236	0.0232
GCC	0.0261	0.0252	0.0259	0.0243	0.0241	0.0242
TAC	0.0238	0.0072	0.0066	0.0071	0.0071	0.0078
CAC	0.0266	0.0102	0.0112	0.0121	0.0122	0.0110
AAC	0.0474	0.0131	0.0126	0.0133	0.0133	0.0135
GAC	0.0065	0.0122	0.0119	0.0125	0.0125	0.0126
TGC	0.0044	0.0079	0.0078	0.0076	0.0075	0.0086
CGC	0.0105	0.0093	0.0104	0.0111	0.0111	0.0101
AGC	0.0174	0.0119	0.0137	0.0136	0.0134	0.0123
GGC	0.0196	0.0108	0.0103	0.0106	0.0106	0.0112
TTA	0.0173	0.0200	0.0211	0.0208	0.0206	0.0200
CTA	0.0918	0.0200	0.0214	0.0209	0.0206	0.0200
ATA	0.0523	0.0172	0.0176	0.0188	0.0188	0.0184
GTA	0.0174	0.0169	0.0153	0.0168	0.0171	0.0180
TCA	0.0218	0.0187	0.0188	0.0185	0.0184	0.0186
CCA	0.0131	0.0196	0.0209	0.0204	0.0202	0.0195
ACA	0.0480	0.0191	0.0191	0.0204	0.0205	0.0203
GCA	0.0174	0.0157	0.0145	0.0159	0.0161	0.0168
CAA	0.0196	0.0088	0.0091	0.0078	0.0077	0.0082
AAA	0.0219	0.0116	0.0120	0.0116	0.0115	0.0108
GAA	0.0195	0.0103	0.0086	0.0087	0.0088	0.0096
TGA	0.0261	0.0124	0.0126	0.0116	0.0116	0.0119
CGA	0.0087	0.0107	0.0110	0.0099	0.0099	0.0102
GGA	0.0107	0.0095	0.0084	0.0085	0.0086	0.0088
TTG	0.0022	0.0200	0.0211	0.0208	0.0206	0.0200
CTG	0.0086	0.0200	0.0214	0.0209	0.0206	0.0200
ATG	0.0065	0.0172	0.0176	0.0188	0.0188	0.0184
GTG	0.0022	0.0169	0.0153	0.0168	0.0171	0.0180
TCG	0.0022	0.0187	0.0188	0.0185	0.0184	0.0186
CCG	0.0000	0.0196	0.0209	0.0204	0.0202	0.0195
ACG	0.0021	0.0191	0.0191	0.0204	0.0205	0.0203
GCG	0.0000	0.0157	0.0145	0.0159	0.0161	0.0168
CAG	0.0022	0.0095	0.0097	0.0084	0.0083	0.0090
AAG	0.0022	0.0116	0.0120	0.0116	0.0115	0.0108
GAG	0.0000	0.0103	0.0086	0.0087	0.0088	0.0096
TGG	0.0022	0.0124	0.0126	0.0116	0.0116	0.0119
CGG	0.0000	0.0107	0.0110	0.0099	0.0099	0.0102
GGG	0.0046	0.0095	0.0084	0.0085	0.0086	0.0088

Tabela 7.5: Probabilidades Estimadas dos Modelos Binomiais Multivariados com Função de Ligação Log-Log Complementar

Códon	Obs.	Modelo Independente	Modelo Ig. Preditivo	Estrutura Markoviana	Modelo Aditivo	Modelo Bahadur
TTT	0.0196	0.0309	0.0295	0.0306	0.0298	0.0297
CTT	0.0219	0.0338	0.0320	0.0346	0.0329	0.0326
ATT	0.0349	0.0250	0.0266	0.0238	0.0249	0.0255
GTT	0.0000	0.0240	0.0254	0.0226	0.0237	0.0244
TCT	0.0109	0.0290	0.0280	0.0280	0.0278	0.0278
CCT	0.0055	0.0227	0.0196	0.0216	0.0202	0.0214
ACT	0.0174	0.0199	0.0203	0.0178	0.0187	0.0202
GCT	0.0131	0.0261	0.0273	0.0250	0.0258	0.0266
TAT	0.0045	0.0080	0.0075	0.0079	0.0075	0.0088
CAT	0.0022	0.0086	0.0091	0.0101	0.0099	0.0094
AAT	0.0027	0.0148	0.0157	0.0167	0.0171	0.0145
GAT	0.0000	0.0097	0.0089	0.0098	0.0096	0.0095
TGT	0.0022	0.0103	0.0094	0.0100	0.0092	0.0107
CGT	0.0000	0.0098	0.0111	0.0115	0.0117	0.0107
AGT	0.0044	0.0122	0.0138	0.0143	0.0154	0.0120
GGT	0.0022	0.0101	0.0095	0.0101	0.0101	0.0098
TTC	0.0240	0.0309	0.0295	0.0306	0.0298	0.0297
CTC	0.0674	0.0338	0.0320	0.0346	0.0329	0.0326
ATC	0.0501	0.0250	0.0266	0.0238	0.0249	0.0255
GTC	0.0087	0.0240	0.0254	0.0226	0.0237	0.0244
TCC	0.0370	0.0290	0.0280	0.0280	0.0278	0.0278
CCC	0.0316	0.0227	0.0196	0.0216	0.0202	0.0214
ACC	0.0371	0.0199	0.0203	0.0178	0.0187	0.0202
GCC	0.0261	0.0261	0.0273	0.0250	0.0258	0.0266
TAC	0.0238	0.0080	0.0075	0.0078	0.0075	0.0088
CAC	0.0266	0.0086	0.0091	0.0101	0.0099	0.0094
AAC	0.0474	0.0148	0.0157	0.0167	0.0171	0.0145
GAC	0.0065	0.0097	0.0089	0.0098	0.0096	0.0095
TGC	0.0044	0.0103	0.0094	0.0100	0.0092	0.0107
CGC	0.0105	0.0098	0.0111	0.0115	0.0117	0.0107
AGC	0.0174	0.0122	0.0138	0.0143	0.0154	0.0120
GGC	0.0196	0.0101	0.0095	0.0101	0.0101	0.0098
TTA	0.0173	0.0178	0.0186	0.0186	0.0196	0.0193
CTA	0.0918	0.0178	0.0187	0.0187	0.0198	0.0192
ATA	0.0523	0.0181	0.0183	0.0195	0.0189	0.0174
GTA	0.0174	0.0184	0.0178	0.0191	0.0181	0.0177
TCA	0.0218	0.0176	0.0184	0.0183	0.0193	0.0191
CCA	0.0131	0.0165	0.0169	0.0172	0.0180	0.0181
ACA	0.0480	0.0240	0.0251	0.0263	0.0253	0.0235
GCA	0.0174	0.0134	0.0120	0.0132	0.0123	0.0127
CAA	0.0196	0.0116	0.0128	0.0101	0.0107	0.0104
AAA	0.0219	0.0128	0.0124	0.0124	0.0122	0.0132
GAA	0.0195	0.0112	0.0095	0.0097	0.0090	0.0117
TGA	0.0261	0.0089	0.0096	0.0081	0.0083	0.0076
CGA	0.0087	0.0118	0.0128	0.0103	0.0108	0.0106
GGA	0.0107	0.0051	0.0036	0.0039	0.0035	0.0056
TTG	0.0022	0.0178	0.0186	0.0186	0.0196	0.0193
CTG	0.0086	0.0178	0.0187	0.0187	0.0198	0.0192
ATG	0.0065	0.0181	0.0183	0.0195	0.0189	0.0174
GTG	0.0022	0.0184	0.0178	0.0191	0.0181	0.0177
TCG	0.0022	0.0176	0.0184	0.0183	0.0193	0.0191
CCG	0.0000	0.0165	0.0169	0.0172	0.0180	0.0181
ACG	0.0021	0.0240	0.0251	0.0263	0.0253	0.0235
GCG	0.0000	0.0134	0.0120	0.0132	0.0123	0.0127
CAG	0.0022	0.0122	0.0126	0.0104	0.0108	0.0112
AAG	0.0022	0.0128	0.0124	0.0124	0.0122	0.0132
GAG	0.0000	0.0112	0.0095	0.0097	0.0090	0.0117
TGG	0.0022	0.0089	0.0096	0.0081	0.0083	0.0076
CGG	0.0000	0.0118	0.0128	0.0103	0.0108	0.0106
GGG	0.0046	0.0051	0.0036	0.0039	0.0035	0.0056

Tabela 7.6: Probabilidades Estimadas dos Modelos Multinomiais Multivariados Logísticos Regressivos

Códon	Obs.	Modelo Independente	Modelo Ig. Preditivo	Estrutura Markoviana	Modelo Aditivo
TTT	0.0196	0.0118	0.0115	0.0171	0.0130
CTT	0.0219	0.0164	0.0269	0.0291	0.0233
ATT	0.0349	0.0215	0.0159	0.0255	0.0349
GTT	0.0000	0.0087	0.0023	0.0052	0.0047
TCT	0.0109	0.0085	0.0115	0.0168	0.0141
CCT	0.0055	0.0123	0.0177	0.0068	0.0052
ACT	0.0174	0.0143	0.0172	0.0136	0.0190
GCT	0.0131	0.0051	0.0043	0.0080	0.0081
TAT	0.0045	0.0065	0.0072	0.0058	0.0043
CAT	0.0022	0.0086	0.0081	0.0007	0.0005
AAT	0.0027	0.0093	0.0076	0.0023	0.0035
GAT	0.0000	0.0034	0.0019	0.0013	0.0013
TGT	0.0022	0.0015	0.0022	0.0071	0.0058
CGT	0.0000	0.0059	0.0033	0.0005	0.0004
AGT	0.0044	0.0059	0.0028	0.0017	0.0025
GGT	0.0022	0.0023	0.0007	0.0008	0.0008
TTC	0.0240	0.0391	0.0378	0.0268	0.0282
CTC	0.0674	0.0555	0.0660	0.0760	0.0791
ATC	0.0501	0.0645	0.0616	0.0531	0.0486
GTC	0.0087	0.0247	0.0162	0.0120	0.0123
TCC	0.0370	0.0280	0.0226	0.0220	0.0236
CCC	0.0316	0.0366	0.0254	0.0199	0.0194
ACC	0.0371	0.0348	0.0410	0.0543	0.0532
GCC	0.0261	0.0173	0.0230	0.0200	0.0198
TAC	0.0238	0.0193	0.0188	0.0300	0.0322
CAC	0.0266	0.0200	0.0173	0.0223	0.0233
AAC	0.0474	0.0282	0.0302	0.0319	0.0310
GAC	0.0065	0.0124	0.0161	0.0135	0.0127
TGC	0.0044	0.0097	0.0136	0.0075	0.0055
CGC	0.0105	0.0147	0.0124	0.0128	0.0139
AGC	0.0174	0.0246	0.0243	0.0186	0.0177
GGC	0.0196	0.0084	0.0120	0.0170	0.0177
TTA	0.0173	0.0311	0.0305	0.0208	0.0219
CTA	0.0918	0.0535	0.0648	0.0779	0.0810
ATA	0.0523	0.0605	0.0596	0.0564	0.0513
GTA	0.0174	0.0236	0.0155	0.0113	0.0116
TCA	0.0218	0.0258	0.0220	0.0225	0.0244
CCA	0.0131	0.0352	0.0250	0.0204	0.0199
ACA	0.0480	0.0372	0.0409	0.0476	0.0459
GCA	0.0174	0.0168	0.0228	0.0210	0.0208
CAA	0.0196	0.0230	0.0190	0.0203	0.0204
AAA	0.0219	0.0312	0.0314	0.0263	0.0253
GAA	0.0195	0.0111	0.0158	0.0182	0.0186
TGA	0.0261	0.0116	0.0138	0.0189	0.0202
CGA	0.0087	0.0159	0.0128	0.0102	0.0109
GGA	0.0107	0.0087	0.0117	0.0135	0.0136
TTG	0.0022	0.0028	0.0028	0.0019	0.0020
CTG	0.0086	0.0049	0.0059	0.0070	0.0073
ATG	0.0065	0.0055	0.0054	0.0051	0.0046
GTG	0.0022	0.0022	0.0014	0.0010	0.0010
TCG	0.0022	0.0024	0.0020	0.0020	0.0022
CCG	0.0000	0.0032	0.0023	0.0018	0.0018
ACG	0.0021	0.0034	0.0037	0.0043	0.0041
GCG	0.0000	0.0015	0.0021	0.0019	0.0019
CAG	0.0022	0.0019	0.0016	0.0020	0.0021
AAG	0.0022	0.0028	0.0029	0.0024	0.0023
GAG	0.0000	0.0010	0.0014	0.0017	0.0017
TGG	0.0022	0.0011	0.0013	0.0017	0.0018
CGG	0.0000	0.0015	0.0012	0.0009	0.0010
GGG	0.0046	0.0008	0.0011	0.0012	0.0012

Tabela 7.7: Probabilidades Estimadas dos Modelos Multinomiais Multivariados Baseados na Representação de Bahadur

Códon	Obs.	Locação	Transição	Semi-Locação e Transição	Locação e Transição
TTT	0.0196	0.0135	0.0120	0.0143	0.0138
CTT	0.0219	0.0188	0.0242	0.0184	0.0224
ATT	0.0349	0.0244	0.0202	0.0224	0.0254
GTT	0.0000	0.0099	0.0008	0.0000	0.0049
TCT	0.0109	0.0097	0.0101	0.0136	0.0106
CCT	0.0055	0.0108	0.0169	0.0123	0.0067
ACT	0.0174	0.0123	0.0198	0.0271	0.0174
GCT	0.0131	0.0041	0.0045	0.0051	0.0052
TAT	0.0045	0.0073	0.0067	0.0073	0.0050
CAT	0.0022	0.0069	0.0091	0.0024	0.0029
AAT	0.0027	0.0079	0.0076	0.0095	0.0049
GAT	0.0000	0.0026	0.0026	0.0010	0.0034
TGT	0.0022	0.0017	0.0009	0.0015	0.0014
CGT	0.0000	0.0045	0.0031	0.0000	0.0005
AGT	0.0044	0.0050	0.0024	0.0038	0.0054
GGT	0.0022	0.0016	0.0007	0.0013	0.0019
TTC	0.0240	0.0331	0.0371	0.0330	0.0286
CTC	0.0674	0.0567	0.0650	0.0738	0.0763
ATC	0.0501	0.0656	0.0563	0.0497	0.0493
GTC	0.0087	0.0268	0.0154	0.0135	0.0076
TCC	0.0370	0.0297	0.0244	0.0215	0.0269
CCC	0.0316	0.0363	0.0300	0.0277	0.0267
ACC	0.0371	0.0346	0.0372	0.0404	0.0342
GCC	0.0261	0.0167	0.0200	0.0185	0.0157
TAC	0.0238	0.0212	0.0243	0.0242	0.0206
CAC	0.0266	0.0196	0.0210	0.0211	0.0230
AAC	0.0474	0.0279	0.0291	0.0351	0.0233
GAC	0.0065	0.0118	0.0196	0.0158	0.0179
TGC	0.0044	0.0106	0.0101	0.0114	0.0108
CGC	0.0105	0.0144	0.0124	0.0125	0.0130
AGC	0.0174	0.0244	0.0221	0.0224	0.0282
GGC	0.0196	0.0079	0.0116	0.0138	0.0104
TTA	0.0173	0.0262	0.0322	0.0308	0.0265
CTA	0.0918	0.0545	0.0634	0.0744	0.0667
ATA	0.0523	0.0614	0.0651	0.0531	0.0582
GTA	0.0174	0.0256	0.0162	0.0221	0.0182
TCA	0.0218	0.0272	0.0230	0.0217	0.0237
CCA	0.0131	0.0350	0.0268	0.0257	0.0148
ACA	0.0480	0.0370	0.0444	0.0436	0.0374
GCA	0.0174	0.0163	0.0194	0.0232	0.0201
CAA	0.0196	0.0227	0.0149	0.0157	0.0194
AAA	0.0219	0.0310	0.0253	0.0284	0.0263
GAA	0.0195	0.0106	0.0139	0.0141	0.0198
TGA	0.0261	0.0130	0.0133	0.0128	0.0120
CGA	0.0087	0.0156	0.0136	0.0104	0.0077
GGA	0.0107	0.0082	0.0122	0.0151	0.0121
TTG	0.0022	0.0010	0.0036	0.0042	0.0041
CTG	0.0086	0.0040	0.0046	0.0081	0.0078
ATG	0.0065	0.0045	0.0056	0.0063	0.0076
GTG	0.0022	0.0022	0.0033	0.0012	0.0009
TCG	0.0022	0.0022	0.0015	0.0009	0.0019
CCG	0.0000	0.0038	0.0000	0.0000	0.0000
ACG	0.0021	0.0040	0.0021	0.0014	0.0021
GCG	0.0000	0.0019	0.0023	0.0000	0.0002
CAG	0.0022	0.0024	0.0005	0.0007	0.0016
AAG	0.0022	0.0034	0.0020	0.0019	0.0028
GAG	0.0000	0.0013	0.0021	0.0004	0.0012
TGG	0.0022	0.0012	0.0021	0.0028	0.0012
CGG	0.0000	0.0019	0.0018	0.0030	0.0004
GGG	0.0046	0.0010	0.0023	0.0019	0.0005

# Apêndice 3

## Parâmetros Estimados dos Modelos

Tabela 7.8: Parâmetros Estimados dos Modelos Logístico Regressivos Independentes Binomiais Multivariados

	Função de Ligação		
	Logito	Probit	Log-Log C.
$\alpha_1$	3.3377	0.2374	2.0525
$\alpha_2$	2.6802	-0.2594	1.5797
$\alpha_3$	2.9428	0.0899	1.7494
$\beta_1$	0.0200	0.0120	0.0133
$\beta_2$	-0.4839	-0.1000	-0.3450
$\beta_3$	0.1044	0.0742	0.0725

Tabela 7.9: Parâmetros Estimados dos Modelos Logístico Regressivos Iguamente Preditivos Binomiais Multivariados

	Função de Ligação		
	Logito	Probit	Log-Log C.
$\alpha_1$	3.8215	0.0493	2.5199
$\alpha_2$	3.3575	-0.3863	2.2098
$\alpha_3$	3.6007	0.0185	2.3471
$\gamma$	-0.2573	-0.1425	-0.2026
$\beta_1$	0.0218	0.0114	0.0145
$\beta_2$	-0.5523	-0.0812	-0.4069
$\beta_3$	0.1928	0.1529	0.1314



Tabela 7.10: Parâmetros Estimados dos Modelos Logístico Regressivos com Estrutura Markoviana de Primeira Ordem Binomiais Multivariados

	<b>Função de Ligação</b>		
	<b>Logito</b>	<b>Probit</b>	<b>Log-Log C.</b>
$\alpha_1$	3.6622	0.0472	2.3608
$\alpha_2$	3.0506	-0.3989	1.9432
$\alpha_3$	3.3967	-0.0202	2.1459
$\gamma_1$	-0.0815	-0.1275	-0.0621
$\gamma_2$	-0.5309	-0.2456	-0.3663
$\beta_1$	0.0239	0.0134	0.0160
$\beta_2$	-0.5470	-0.0919	-0.3982
$\beta_3$	0.2060	0.1426	0.1289

Tabela 7.11: Parâmetros Estimados dos Modelos Logístico Regressivos Aditivos Binomiais Multivariados

	<b>Função de Ligação</b>		
	<b>Logito</b>	<b>Probit</b>	<b>Log-Log C.</b>
$\alpha_1$	3.7223	0.0538	2.4734
$\alpha_2$	3.1335	-0.3909	2.0830
$\alpha_3$	3.5624	-0.0266	2.3418
$\gamma_1$	-0.1020	-0.1271	-0.0830
$\gamma_2$	-0.2153	0.0259	-0.1630
$\gamma_3$	-0.5650	-0.2423	-0.4104
$\beta_1$	0.0228	0.0137	0.0153
$\beta_2$	-0.5484	-0.0940	-0.4072
$\beta_3$	0.2367	0.1363	0.1503

Tabela 7.12: Parâmetros Estimados dos Modelos Baseados na Representação de Bahadur Binomiais Multivariados

	<b>Função de Ligação</b>		
	<b>Logito</b>	<b>Probit</b>	<b>Log-Log C.</b>
$\alpha_1$	3.3377	0.2374	2.0525
$\alpha_2$	2.6802	-0.2594	1.5797
$\alpha_3$	2.9428	0.0899	1.7494
$\beta_1$	0.0200	0.0120	0.0133
$\beta_2$	-0.4839	-0.1000	-0.3450
$\beta_3$	0.1044	0.0742	0.0725
$\rho_{12}$	0.0064	-0.0099	0.0077
$\rho_{13}$	-0.0088	0.0199	-0.0080
$\rho_{23}$	-0.0356	-0.0436	-0.0265
$\rho_{123}$	0.0611	-0.0072	0.0562

Tabela 7.13: Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Independente

$\alpha_{11}$	=	-1.0156	$\alpha_{21}$	=	-1.7101	$\alpha_{31}$	=	-0.1736	$\beta_1$	=	-0.0062
$\alpha_{12}$	=	-0.8952	$\alpha_{22}$	=	-2.1042	$\alpha_{32}$	=	-0.2122	$\beta_2$	=	0.1801
$\alpha_{13}$	=	-1.8673	$\alpha_{23}$	=	-2.4727	$\alpha_{33}$	=	-2.6073	$\beta_3$	=	0.0836

Tabela 7.14: Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Igualmente Preditivo

$\alpha_{11}$	=	-0.0734	$\alpha_{31}$	=	0.6978	$\beta_1$	=	-0.0101
$\alpha_{12}$	=	0.0836	$\alpha_{32}$	=	0.6805	$\beta_2$	=	0.1116
$\alpha_{13}$	=	-0.8699	$\alpha_{33}$	=	-1.7174	$\beta_3$	=	-0.0412
$\alpha_{21}$	=	-0.9565	$\gamma_1$	=	-0.4149			
$\alpha_{22}$	=	-1.2782	$\gamma_2$	=	0.1849			
$\alpha_{23}$	=	-1.6346	$\gamma_3$	=	0.8269			

Tabela 7.15: Parâmetros Estimados do Modelo Logístico Regressivo Multinomial Estrutura Markoviana de Primeira Ordem

$\alpha_{11}$	=	3.0029	$\alpha_{31}$	=	3.1517	$\gamma_{21}$	=	0.2102
$\alpha_{12}$	=	3.3215	$\alpha_{32}$	=	3.1754	$\gamma_{22}$	=	2.6663
$\alpha_{13}$	=	2.4031	$\alpha_{33}$	=	0.7702	$\gamma_{23}$	=	2.9711
$\alpha_{21}$	=	2.3889	$\gamma_{11}$	=	-1.4894	$\beta_1$	=	-0.0249
$\alpha_{22}$	=	2.4021	$\gamma_{12}$	=	-0.2139	$\beta_2$	=	-0.1320
$\alpha_{23}$	=	2.3435	$\gamma_{13}$	=	0.4623	$\beta_3$	=	-0.4651

Tabela 7.16: Parâmetros Estimados do Modelo Baseado na Representação de Bahadur de Dependência de Locação

$\alpha_{11}$	=	-1.0156	$\alpha_{31}$	=	-0.1736	$\beta_1$	=	-0.0062
$\alpha_{12}$	=	-0.8952	$\alpha_{32}$	=	-0.2122	$\beta_2$	=	0.1801
$\alpha_{13}$	=	-1.8673	$\alpha_{33}$	=	-2.6073	$\beta_3$	=	0.0836
$\alpha_{21}$	=	-1.7101	$\rho_{12}$	=	-0.0500			
$\alpha_{22}$	=	-2.1042	$\rho_{13}$	=	0.0395			
$\alpha_{23}$	=	-2.4727	$\rho_{23}$	=	0.0499			



# Apêndice 4

## Forma de Jordan

A teoria apresentada a seguir pode ser encontrada em [Hoffman e Kunze \(1971\)](#), e foi utilizada no cálculo das matrizes para análises de diagnósticos.

Seja uma matriz  $\mathbf{A}$   $n \times n$ , com  $n$  autovalores distintos. Essa matriz é dita diagonalizável se existe uma matriz  $\mathbf{S}$ , cujas colunas são formadas pelos autovetores da matriz  $\mathbf{A}$ , tal que,

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}, \quad (7.1)$$

em que  $\mathbf{D}$  é uma matriz diagonal com os autovalores da matriz  $\mathbf{A}$ .

Assim,

$$\mathbf{A} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1} \Rightarrow \mathbf{A}^{1/2} = \mathbf{S}\mathbf{D}^{1/2}\mathbf{S}^{-1}, \quad (7.2)$$

de forma que  $\mathbf{D}^{1/2}$  é uma matriz diagonal cujos elementos são a raiz quadrada dos elementos da diagonal de  $\mathbf{D}$ .

Quando uma matriz  $\mathbf{A}$  não é diagonalizável, ou seja, se tem  $p < n$  autovalores, uma solução é usar a forma de Jordan. Nesse caso, seja  $\mathbf{J}$  uma matriz bloco-diagonal, na forma,

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & \\ & \ddots & \\ & & \mathbf{J}_p \end{bmatrix} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}, \quad (7.3)$$

em que cada bloco  $\mathbf{J}_i$  é uma matriz quadrada formada pelos autovalores de  $\mathbf{A}$ , com

dimensão igual ao número de vezes que o autovalor se repete, da seguinte maneira,

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}. \quad (7.4)$$

A matriz  $\mathbf{P}$  é obtida solucionando o sistema  $\mathbf{AP} = \mathbf{PJ}$ .

Finalmente,

$$\mathbf{A}^{1/2} = \mathbf{PJ}^{1/2}\mathbf{P}^{-1}. \quad (7.5)$$