**Universidade de São Paulo**
**Escola Superior de Agricultura "Luiz de Queiroz"**

# Função da probabilidade da seleção do recurso (RSPF ) na seleção de habitat usando modelos de escolha discreta

**Sandra Vergara Cardozo**

Tese apresentada para obtenção do título de Doutor em Agronomia. Área de Concentração: Estatística e Experimentação Agronômica

**Piracicaba**
**2009**

**Sandra Vergara Cardozo**
**Licenciada em Matemática e Física**

# Função da probabilidade da seleção do recurso (RSPF ) na seleção de habitat usando modelos de escolha discreta

**Orientador:**

**Prof. Dr. Carlos Tadeu dos Santos Dias**

Tese apresentada para obtenção do título de Doutor em Agronomia. Área de Concentração:
Estatística e Experimentação Agronômica

**Piracicaba**

**2009**

**Dados Internacionais de Catalogação na Publicação**
**DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP**

Vergara Cardozo, Sandra
Função da probabilidade da seleção do recurso (RSPF) na seleção de habitat usando modelos de escolha discreta / Sandra Vergara Cardozo. - - Piracicaba, 2009.
86 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura Luiz de Queiroz, 2009.
Bibliografia.

1. Análise de regressão e correlação 2. Boostrap Jackniffe re-amostragem 3. Ecologia animal - - Métodos estatísticos 4. Habitat - Seleção 5. Inferência em população animal 6. Modelos matemáticos 7. Probabilidade aplicada I. Título

CDD 591.51

V494f

# DEDICATÓRIA

*À*
**Deus e Jesús**
*por me dar forças e guiar meus passos.*

*Ao meu esposo,*
**Bryan Frederick**,
*o imenso amor, a suprema dedicação, o apoio incondicional, na batalha de doutorado,*
*Á minha mãe* **Maria Emma** *pela dedicação durante toda sua vida,*
*Á* **Amelia Edwards** *pelo apoio incondicional,*
*Á minha tia avó* **Saturia** *por estar sempre presente nos momentos que mais precisei* **(in memoriam).**

*A minha filha Carolina, a pequena estatística, sempre alegrando nossa vida.*

*Amo muito vocês!*

*Minha eterna gratidão...*

# AGRADECIMENTOS

À minha família e a família do Bryan pelo incentivo carinho e apoio.

Ao meu orientador Prof. Dr. Carlos Tadeu dos Santos, Elvina, Victor e Laurinha pelo conhecimento compartilhado e apoio, os ensinamentos, as sugestões, as discussões, a paciência e amizade, tornando possível a realização deste trabalho.

A Western EcoSystems Technology, Inc, por fornecer a base de dados.

Aos Professores Dr. Bryan Manly, Dr. Chandra Bhat e Dr. Daly Zachary pelas valiosas sugestões os, ensinamentos e a ajuda.

A Prof. Liciana Silva, ao Prof. José Claudio Faria e à Prof. Sônia Maria De Stefano Piedade, pelas valiosas contribuições em meu exame de qualificação.

Aos professores do departamento de Ciências Exatas da ESALQ/USP, Dra Clarice Garcia Borges Démetrio, Dr. Décio Barbin, Dra Roseli Aparecida Leandro, Dra. Sônia Maria De Stefano Piedade, Dr. Silvio Sandoval Zocchi, Dr. César Gonçalves de Lima, Dr. Gerson Barreto Mourão, e Dr. Edwin Moises Marcos Ortega pela amizade, e transmissão de conhecimentos.

A Solange de Assis, Jorge Wiendl, Expedita de Acevedo, Eduardo Bonilha, Luciane Brajão e aos funcionários do departamento de Ciências Exatas ESALQ/USP que sempre estavam prontos para um socorro, uma palavra amiga e dessa maneira propiciaram condições para a realização deste trabalho.

Aos amigos Ramiro Ruíz e Edna Reis pelas sugestões, amizade e troca de conhecimentos.

Aos amigos virtuais dos grupos Biquinho e listas do SAS, pelas sugestões no Latex e SAS.

Aos amigos Melissa Lombardi, Renato Silva, Vanderly Janeiro e o Dr. Silvano Cesar da Costa pelas trocas de conhecimentos de Latex e a amizade.

A Silvio Douglas Dias Bacheta e Vilma Aparecida Sarto Zeferino, pelo excelente serviço de empréstimo entre bibliotecas e demais funcionários das bibliotecas.

Á Capes o suporte financeiro.

À Beatriz Helena Giongo, a correção e a revisão das normas.

Aos colegas de pós-graduação, em especial a, Ana Maria Souza, Geneville Bergano, Idemauro Lara, David Miguelluti, Elisabeth Toledo, Mauricio Mota, José Fogo, Andréia Meyer, Wilson Oliveira, Édila de Souza, Afrânio Vieira, Fernanda Rizzato, Hélio Rubens, Joseane Padilla, Juliana Fachini, Pâmela Piovessam, Alexandre Silva, Júlio Pereira, Luciana Carvalho, Osmar Macedo, Ana Carolina Alexandrino, Osmar Macedo e Alexandre Barbosa (*in memoriam*). Pela ajuda, apoio, compreensão e laços de amizade, em fim a todos os colegas de mestrado e doutorado do Curso de Estatística e Experimentação Agronômica pela amizade e bom convívio.

Aos amigos, Yolanda Borgues, Giovana Silva, Renata Alcarde, Tais Vollstedt, Sandra Maria Marcellino, Sandra Montenegro, Maria Sueli Bonini, Gladys Giraldo, Jairo Jimenez Rueda, Rene Lopera e família, Bia Biscate e família , Patricia Sevilla e Ana Paula Soares e família, por estarem ao meu lado nos momentos que mais precisei.

Aos colegas e amigos das Universidades Nacional de Bogotá, Luis Alberto López e família e da Universidade del Tolima, Alfonso Sanchez e família, Jairo Alfonso Clavijo, Luz Stella Torres, Miguel Angel Quimbayo e família pelo apoio, o incentivo e a confiança depositada.

A todos, que de forma direta o indireta colaboraram para a realização deste trabalho.

“ There is always going to be an element of doubt, as one is extrapolating into areas one doesn’t know about. But what EVT is doing is making the best use of whatever data you have about extreme phenomena.”

(Richard Smith)

# SUMÁRIO

# RESUMO

## Função da probabilidade da seleção do recurso (RSPF ) na seleção de habitat usando modelos de escolha discreta

Em ecologia, o comportamento dos animais é freqüentemente estudado para entender melhor suas preferências por diferentes tipos de alimento e habitat. O presente trabalho esta relacionado a este tópico, dividindo-se em três capítulos. O primeiro capitulo refere-se à estimação da função da probabilidade da seleção de recurso (RSPF) comparado com um modelo de escolha discreta (DCM) com uma escolha, usando as estatísticas qui-quadrado para obter as estimativas. As melhores estimativas foram obtidas pelo método DCM com uma escolha. No entanto, os animais não fazem a sua seleção baseados apenas em uma escolha. Com RSPF, as estimativas de máxima verossimilhança, usadas pela regressão logística ainda não atingiram os objetivos, já que os animais têm mais de uma escolha. R e o software Minitab e a linguagem de programação Fortran foram usados para obter os resultados deste capítulo. No segundo capítulo discutimos mais a verossimilhança do primeiro capítulo. Uma nova verossimilhança para a RSPF é apresentada, a qual considera as unidades usadas e não usadas, e métodos de bootstrapping paramétrico e não paramétrico são usados para estudar o viés e a variância dos estimadores dos parâmetros, usando o programa FORTRAN para obter os resultados. No terceiro capítulo, a nova verossimilhança apresentada no capítulo 2 é usada com um modelo de escolha discreta, para resolver parte do problema apresentado no primeiro capítulo. A estrutura de encaixe é proposta para modelar a seleção de habitat de 28 corujas manchadas (Strix occidentalis), assim como a uma generalização do modelo logit encaixado, usando a maximização da utilidade aleatória e a RSPF aleatória. Métodos de otimização numérica, e o sistema computacional SAS, são usados para estimar os parâmetros de estrutura de encaixe.

Palavras-chave: Função da probabilidade da seleção do recurso (RSPF ); Função da seleção do recurso (RSF ); Modelo de escolha discreta (DCM ); Bootstrapping paramétrico e não paramétrico; Modelo logit encaixado generalizado (GNL ); Máxima utilidade aleatória

# ABSTRACT

## Resource of selection probability function (RSPF ) the habitat selection using discrete choice models (DCM )

In ecology, the behavior of animals is often studied to better understand their preferences for different types of habitat and food. The present work is concerned with this topic. It is divided into three chapters. The first concerns the estimation of a resource selection probability function (RSPF) compared with a discrete choice model (DCM) using chi-squared to obtain estimates. The best estimates were obtained by the DCM method. Nevertheless, animals were not selected based on choice alone. With RSPF, the maximum likelihood estimates used with the logistic regression still did not reach the objectives, since the animals have more than one choice. R and Minitab software and the FORTRAN programming language were used for the computations in this chapter. The second chapter discusses further the likelihood presented in the first chapter. A new likelihood for a RSPF is presented, which takes into account the units used and not used, and parametric and non-parametric bootstrapping are employed to study the bias and variance of parameter estimators, using a FORTRAN program for the calculations. In the third chapter, the new likelihood presented in chapter 2, with a discrete choice model is used to resolve a part of the problem presented in the first chapter. A nested structure is proposed for modelling selection by 28 spotted owls (*Strix occidentalis*) as well as a generalized nested logit model using random utility maximization and a random RSPF. Numerical optimization methods and the SAS system were employed to estimate the nested structural parameters.

Keywords: Resource selection probability function (RSPF); Resource selection function (RSF); Discrete choice model (DCM); Parametric and non-parametric bootstrapping; Generalized nested logit model (GNL); Random utility maximization

# LISTA DE SÍMBOLOS

$n$ : n-ésimo grupos de animais na situação de escolha qualitativa, $n = 1, \ldots, N$.

$w(\cdot)$ : A função vetor dos dados observados.

$b$ : número de escolhas de alimentos ou habitat pelo animal ou grupo de animais.

$t$ : O animal ou grupo de animais que define a probabilidade de escolha.

$i$ : A alternativa de escolher um alimento ou característica do habitat $i = 1$.

$j$ : As alternativas de escolher o alimento ou características de habitat $j = 2, \ldots, J$.

$J_n$ : O conjunto de alternativas ou escolha do habitat ou alimento, pois diferentes grupos de animais podem escolher diferentes conjuntos de alternativas em situações de escolha similares.

$\mathbf{s}_n$ : Vetor das características observadas dos $n$ grupos de animais e observadas pelo pesquisador.

$P_{ni}$ : A probabilidade de que $n$ grupos de animais escolham uma alternativa $i$ do conjunto $J_n$, dependendo das características observadas da alternativa $i$ comparada com todas as outras alternativas.

$\mathbf{z}_{ni}$ : Vetor das características das alternativas que são observadas pelo pesquisador, dos $n$ grupos de animais na alternativa $i$.

$\mathbf{z}_{nj}$ : Vetor das características observadas das alternativas $j$, que são observados pelo pesquisador dos $n$ grupos de animais.

$U_{ni}$ : A utilidade dos $n$ grupos de animais obter a alternativa $i$ dentro de $J_n$ (e similarmente para cada outra alternativa dentro de $J_n$).

$\mathbf{r}_a$ : Vetor de todas as características relevantes do animal a, $a = 1, \ldots, A$.

$\mathbf{x}_{ai}$ : Vetor observado de todas as características relevantes da alternativa $i$ como vista pelo animal $a$.

$j\,j'$ : Elasticidade cruzada do par de alternativas $j$ e $j'$.

$d$ : A pessoa.

$k$ : Número de encaixes, Os subconjuntos não sobrepostos, $k = 1, \ldots, K$.

$m$ : As alternativas no encaixe $k$, $m = 1, \ldots, M$.

$l$ : As alternativas no encaixe $l$, $l = 1, \ldots, L$.

$r$ : Conjunto de alternativas selecionadas dentro de um encaixe.

$h$ : O período.

$f$ : f-ésimas derivadas parciais.

$q(k|j)$ : Seleção de *k* alternativas dado que as *j* alternativas foram preferidas.

$k_n$ : O subconjunto selecionado pelos *n* grupos de animais.

$\boldsymbol{\lambda}_k$ : O parâmetro $\boldsymbol{\lambda}_k$ é a medida do grau de independência na utilidade não observada entre as alternativas do encaixe *k*.

$\boldsymbol{\alpha}$ : O vetor de parâmetros a ser estimado que entram na utilidade representativa.

$W_{nk}$ : Constante para todas as alternativas dentro do encaixe, dos *n* grupos de animais dependendo das variáveis que descreve o encaixe *k*.

$M_e$ : O número de plantas presentes na área *e*.

U : Função de utilidade.

$U_{ti}$ : O valor da utilidade, a representação numérica das $t's$ preferências sobre as alternativas dentro de $C_t$, para o recurso da unidade *i* e o animal *t* que define a probabilidade de escolha.

$C_r$ : Conjunto genérico de alternativas que pertencem ao encaixe *r*.

$N_r$ : O conjunto de alternativas incluídas no encaixe *r*.

$\boldsymbol{\alpha}_{j'r}$ : Parâmetro de atribuição dos *n* grupos de animais que caracteriza a parte da alternativa $rj'$ atribuída ao encaixe *r*.

$Y_j$ : Valor de cada alternativa.

# 1 INTRODUÇÃO

A determinação de quais recursos são mais freqüentemente selecionados por uma espécie é de particular interesse, podendo fornecer informações fundamentais sobre a natureza dos animais e sobre as exigências para sua sobrevivência. Assim, quando uma espécie seleciona melhor seus recursos, poderá satisfazer suas exigências vitais com recursos de alta qualidade.

Segundo Rosenzweig (1981), a seleção diferencial dos recursos disponíveis, é um dos principais fatores que permitem a coexistência das espécies. Para sustentar as populações de animais são necessárias quantidades adequadas desses recursos. Cabe aos biólogos identificar os recursos usados pelos animais e avaliar sua disponibilidade no ambiente. Essa necessidade é especialmente crítica para preservar as espécies em risco de extinção, além de poder servir para a exploração sustentável das espécies de interesse econômico.

Supõe-se, freqüentemente, que uma espécie selecionará os melhores recursos para suas exigências de vida, e que um recurso de qualidade mais elevada é, preferencialmente, selecionado em relação a um de baixa qualidade. No entanto, a disponibilidade de vários recursos geralmente não é uniforme na natureza e sua utilização pode depender da sua disponibilidade. De acordo com sua seleção, os recursos disponíveis para uma determinada espécie podem ou não ser utilizados, e é de interesse comparar esses dois tipos de unidades de recurso (usadas *vs* disponíveis). O recurso usado por um animal corresponde àquele que ele seleciona do ambiente em um período de tempo fixo, enquanto que a disponibilidade de um recurso é a quantidade acessível ao animal ou à população durante o mesmo período de tempo. O uso de um recurso é chamado de seletivo, se tal recurso é usado desproporcionalmente à sua disponibilidade.

Existe uma variação considerável na motivação para dirigir os estudos de Seleção do Recurso (RS). Às vezes necessita-se de informações quantitativas sobre as preferências de seleção do recurso para uma determinada espécie que sirvam como indicativo das exigências do recurso a longo prazo para tal espécie.

De acordo com Manly et al. (2002) os estudos quantitativos sobre seleção de recursos por animais se iniciaram há várias décadas, sendo Scott (1920) o primeiro a discutir como esses estudos poderiam ser realizados. Ele dividiu as abundâncias relativas de diferentes tipos de alimento (plâncton), encontrados no estômago de peixes, pela abundância relativa observada em amostras de plâncton coletadas nas mesmas áreas. Este foi o primeiro tipo de aplicação encontrado na literatura sobre a comparação quantitativa de amostras de recursos usados (plâncton no estômago de peixes) com amostras de recursos disponíveis (plâncton).

A busca de uma maneira apropriada de quantificar a RS (alimento ou habitat) pelos animais resultou em um grande número de índices de seleção alternativos que foram propostos a partir da sugestão inicial de Scott (1920). Manly et al. (2002) listam diferentes índices propostos entre 1931 a 1998, e ressaltam que alguns deles, não têm um claro significado biológico.

Com os índices de seleção, surgiram inúmeras propostas para testes de hipóteses de não seleção dos recursos pelos animais, incluíndo a aplicação de métodos padrões tais como: testes de Qui-Quadrado, modelos log-lineares, análise de função discriminante e regressão logística, entre outros; assim como métodos desenhados especificamente para dados de RS. Manly et al. (2002) apresentam uma lista dos principais testes para dados discretos e contínuos propostos entre 1974 e 1998.

Baseado na sua revisão e nas mais recentes propostas de análises, Manly et al. (2002) ressaltam a carên-

cia de uma teoria estatística unificada que mostre como a RS pode ser detectada e medida. Eles argumentam que a falta de tal teoria, fez com que muitos índices e métodos de análises não usuais fossem propostos, dificultando a identificação da melhor aproximação estatística. Os autores sugerem que o conceito da Função da Seleção do Recurso (RSF ) pode ser a base de desenvolvimento dessa teoria.

A definição da RSF , requer que o recurso que está sendo estudado consista em unidades discretas. A população do recurso disponível é então definida como o conjunto destas unidades de recurso disponíveis para o uso do animal ou animais que estão sendo estudados. Por exemplo, se o recurso usado pelo animal é um determinado tipo de inseto, as unidades de recursos são o número de tais insetos disponíveis que podem ser consumidos pelo animal na área onde é examinado o alimento. Neste caso, o estudo da RS poderia estar relacionado a se os animais, para a sua alimentação, preferem esse tipo de inseto.

Se o recurso é o habitat, então é comum dividir a área geográfica que está sendo estudada em parcelas de tamanho fixo, por exemplo de 100m $\times$ 100m. A população de unidades disponíveis é o conjunto de todas as parcelas não sobrepostas na área de estudo. Nesse caso, o estudo da RS poderia estar relacionado a se o animal ou animais que estão sendo estudados tendem a encontrar unidades do recurso com características particulares, por exemplo, locais perto da água.

Em outros casos, supõe-se que as unidades podem ser descritas por *p* variáveis $X_1$, $X_2$,..., $X_p$, que podem ser contínuas (ex.: o comprimento de um inseto, ou a distância do animal objeto de estudo à água), ou representar níveis de um fator (ex.: cor da presa, ou um tipo de vegetação particular). Dada essa situação, Manly et al. (2002) definem a Função de Probabilidade da Seleção do Recurso (RSPF ) como a função de probabilidade de ($X_1$, $X_2$,..., $X_p$) para assumir os valores ($x_1$, $x_2$,..., $x_p$) selecionados em uma área específica durante um determinado período de tempo. Porém, dependendo da natureza dos dados, pode ser impossível obter informação para estimar a função de probabilidade com base nesta definição. Por outro lado, a função da seleção do recurso (RSF ) é definida como qualquer função proporcional à RSPF . Isto é, RSF = $k$ RSPF , para alguma constante positiva $k$.

Usualmente, o mais importante é a probabilidade relativa da seleção para diferentes recursos. Por exemplo, ao se fazer uma avaliação do habitat *versus* espécies em extinção, é suficiente conhecer as probabilidades relativas de uso para estimar a RSF .

Em muitos estudos, a RS para um número de animais individuais é acompanhada por um considerável período de tempo, de tal modo que haja uma quantidade razoável de dados para cada animal. Surge então a pergunta: Como analisar dados com esta estrutura? É provável que os animais individualmente não selecionem os recursos da mesma maneira? Pesquisas feitas no passado consideravam a estimação da RSF para cada animal separadamente, e estimavam a RSF para toda a população de animais como sendo a média dessas RSF individuais.

O método estatístico mais freqüentemente usado para estimar a RSF é a regressão logística, mas modelos de escolha discreta também têm sido utilizados recentemente. A teoria de modelos de escolha discreta tem sido bem desenvolvida para as análises de dados em escolhas humanas (TRAIN, 2003), mas segundo McDonald et al. (2006), eles podem ser aplicados também a escolhas do animal com algumas modificações.

Diferentes métodos podem ser usados para coletar os dados. A disponibilidade dos recursos pode ser avaliada por mapas, digitalização, usando Sistema de Informação Geográfica (GIS), ou amostragem por seleção aleatória dos locais de coleta. Os procedimentos de análise podem ser discretos ou contínuos e são,

geralmente, multivariados. Uma importante decisão no delineamento de estudos da Seleção do Recurso (RS ) é a escolha da área de estudo e seus limites, que pode ter um forte impacto nas análises subseqüentes dos dados, especialmente se as unidades do recurso seguem um padrão agregado. A RS pode ser detectada e medida comparando dois dos três possíveis conjuntos de unidades de recurso (usados, não usados e disponíveis).

De acordo com Manly et al. (2002), usualmente são considerados três protocolos de amostragem:

i) Protocolo A: as unidades do recurso disponíveis são aleatoriamente amostradas ou obtidas por censo e as unidades utilizadas são amostradas aleatoriamente.

ii) Protocolo B: as unidades de recurso disponíveis são aleatoriamente amostradas ou obtidas por censo e as amostras aleatórias de unidades de recurso não utilizadas são usadas.

iii) Protocolo C: as unidades de recurso usadas e não usadas são amostradas independentemente.

Cada um destes protocolos de amostragem, pode ser implementado para cada delineamento amostral. Manly et al. (2002) sugerem três tipos de delineamentos para avaliar a seleção:

i) Delineamento I: as medidas de uso e disponibilidade só consideram o total dos animais na área de estudo, sem a identificação dos indivíduos.

ii) Delineamento II: os animais são identificados individualmente e o uso das unidades do recurso é medido para cada um, mas a disponibilidade do recurso é medida para o total da população.

iii) Delineamento III: os indivíduos são identificados e a disponibilidade é medida para cada animal.

Uma das maneiras mais simples para estimar a RSPF envolve o uso do censo das unidades do recurso usadas e não usadas na população, ajustando-se a função de regressão logística para a probabilidade de uso como uma função das variáveis.

Os modelos de escolha discreta são recomendados como uma aproximação para estimar a RSF , quando a escolha do animal ou grupo de animais envolve diferentes conjuntos de unidades de recursos disponíveis.

Os modelos de escolha discreta são poderosos e flexíveis para estudar a seleção do habitat, pois permitem que a disponibilidade do recurso se altere em cada escolha. McDonald et al. (2006) propuseram uma generalização do modelo de escolha discreta para situações em que múltiplas escolhas são feitas, a partir de um ou mais conjuntos de escolha, e só uma amostra aleatória de cada conjunto está disponível, e quando as escolhas são feitas com reposição ou a ordem temporal da seleção é conhecida, ou desconhecida.

Os modelos de escolha discreta supõem que os recursos selecionados e os conjuntos únicos de unidades do recurso estão disponíveis para cada escolha. Por exemplo, o conjunto de escolhas poderia ser definido como os locais freqüentados dentro de uma pequena área geográfica, a serem usados por um animal. Como o animal está em movimento, o conjunto de escolhas se altera. Assim, apresentam-se a um animal diferentes alimentos A, B e C em um instante de tempo, mas a escolha poderia ser entre os alimentos A, B, e D, em outro instante. Segundo Manly et al. (2002), com um modelo de escolha discreta para a RS , a *i*-ésima escolha é descrita por:

1. o conjunto de escolha de $n_i$ unidades de recurso (habitat ou alimento) que estão disponíveis para serem escolhidas;

2. valores para as variáveis que caracterizam todas as unidades de recurso no conjunto de escolha (ex.: tipo de vegetação, elevação do terreno etc).

O uso de um recurso pode diferir de acordo com sua disponibilidade casual. Se os animais precisarem de uma quantidade particular desse recurso, eles poderiam evitá-lo quando fosse abundante para evitar que se torne escasso. Manly et al. (2002) criticam a suposição dessa seleção, que é independente da disponibilidade, como um tratamento de respostas funcionais que não foi bem resolvido.

Uma pesquisa sobre a seleção de habitat noturno da coruja manchada (*Strix occidentalis*) foi feita para identificar seus locais de preferência na área de Green Diamond Resource Company (NIELSON; McDONALD; LAMPHEAR, 2004). A Função de Seleção do Habitat (HSF ) foi estimada aplicando a teoria proposta por Manly et al. (2002).

Em economia, os modelos de escolha discreta descrevem (as decisões do fabricante) entre diferentes alternativas usando métodos estatísticos para examinar as escolhas de consumidores, famílias etc. Muitas dessas análises são feitas através de simulações. No entanto, o conteúdo dessa teoria ainda não foi verificado quanto à possibilidade de ser usada em modelos de escolha discreta para animais, visto que as condições de escolha de mercadologia para as pessoas não são muitas e os animais podem ter milhões de escolhas de habitat.

Sendo assim, os três capítulos que compõem este trabalho tem como objetivo apresentar o modelo de escolha discreta para identificar as preferências da seleção de habitat de uma espécie animal. Isto foi feito usando dados de um estudo sobre a coruja manchada (*Strix occidentalis*) (NIELSON; McDONALD; LAMPHEAR, 2004), assim como dados simulados.

O primeiro capítulo apresenta uma comparação entre a regressão logística e o modelo de escolha discreta (DCM ) com uma escolha na identificação dos recursos críticos pela população de animais e na predição da ocorrência da espécie. A melhor estimativa foi obtida com o DCM para uma escolha da coruja.

No segundo capítulo, é apresentada uma nova proposta para a estimação da função de verossimilhança da probabilidade dos dados observados, para superar os problemas de viés quando o tamanho amostral é pequeno. Para tal fim foram utilizados três métodos de reamostragem: Bootstrap paramétrico, Bootstrap não paramétrico, e Bootstrap paramétrico com modificações.

No terceiro capítulo, apresenta-se o modelo logit encaixado generalizado como uma proposta para selecionar as preferências de habitat de cada animal e assim generalizar para os grupos de animais, pois estes escolhem uma alternativa que fornece a máxima utilidade, levando em consideração que as escolhas da coruja manchada (*Strix occidentalis*) são múltiplas e independentes.

**Referências**

NIELSON, R.; McDONALD, T.; LAMPHEAR, D. Northern spotted owl nighttime site selection model. **Report Western EcoSystems Technology**, Cheyenne, v. 10, p. 38-43, 2004.

MANLY, B.F.J.; McDONALD, L.L.; THOMAS, D.L.; McDONALD, T.L.; ERICKSON, W.P. **Resource selection by animals.** 2nd ed. London: Kluwer Academic Publishers, 2002. 221 p.

McDONALD, T.; MANLY, B.F.J.; NIELSON, R.M.; DILLER, L.V. Discrete-choice modeling in wildlife studies exemplified by northern spotted owl nighttime habitat selection. **Journal of Wildlife Management**, Cheyenne, v. 70, p. 375-383, 2006.

ROSENZWEIGH, M.L. A theory of habitat selection. **Ecology**, Brooklyn, v. 62, p. 327-335, 1981.

SCOTT, A. Food of port erin mackerel in 1919. **Proceedings and Transactions of the Liverpool Biological Society**, Liverpool, v. 34, p. 107-11, 1920.

TRAIN. **Discrete choice methods with simulation**. Cambridge: Cambridge University Press, 2003. 334 p.

# 2 COMPARAÇÃO DOS MÉTODOS DE REGRESSÃO LOGÍSTICA E MODELO DE ESCOLHA DISCRETA NA SELEÇÃO DO HABITAT DA CORUJA MANCHADA (*Strix occidentalis*)


## Resumo

Baseado na sua revisão e nas mais recentes propostas de análises, Manly et al. (2002) argumentam que falta uma teoria estatística unificada, que mostre como a seleção de recursos pode ser detectada e medida. Os autores sugerem que o conceito da RSF pode ser a base de desenvolvimento da teoria. Os autores sugerem ainda a revisão de Modelos de Escolha Discreta (DCM). A definição da Função da Seleção do Recurso (RSF), requer que o recurso que está sendo estudado consista em unidades discretas. O método estatístico mais freqüentemente utilizado para estimar a RSF é a regressão logística, mas DCM recentemente desenvolvidos também estão sendo utilizados, como uma aproximação para estimar a RSF, quando a escolha do animal ou de um grupo de animais envolve diferentes conjuntos de unidades de recursos disponíveis. A teoria do DCM tem sido bem desenvolvida, para as análises de dados em escolhas de produtos pelos humanos (TRAIN, 2003), mas podem ser aplicadas também às escolhas de habitat pelos animais, com algumas modificações (McDONALD et al., 2006). O objetivo do presente trabalho foi comparar as estimativas das funções da seleção do recurso aplicando regressão logística e DCM a dados provenientes de um estudo de seleção de habitat da coruja manchada (*Strix occidentalis*) no noroeste dos Estados Unidos. Programas computacionais foram desenvolvidos para atingir tal objetivo.


Palavras-chave: Regressão logística; Função da seleção do recurso (RSF); Modelos de escolha discreta (DCM)


## Abstract

Based on their review of early analyses of data on resource selection by animals as well as on more recent suggestions, Manly et al. (2002) argued that there is a lack of a unified statistical theory that shows how resource selection can be detected and measured. They suggest that the concept of RSF can be the base for the development of the needed general theory. Manly et al. (2002) also suggest the use of discrete choice models (DCM) as an approximation to estimate the RSF when the choice of an animal or a group of animals involve different sets of available resource units. The definition of the resource selection function (RSF) requires that the resource which is being studied consists of discrete units. The statistical method often used to estimate the RSF is the logistic regression but DCM can also be used. The theory of DCM has been well developed for the analysis of data sets involving choices of products by humans (TRAIN, 2003), but it can also be applied for the choice of habitat by animals, with some modifications, (McDONALD et al., 2006). The objective of this work was to compare the estimates of the resource selection function obtained by applying the logistic regression and the DCM to a data set from one study on habitat selection of the spotted owl (*Strix occidentalis*) in the north west of the United States.


Keywords: Logistic regression; Resource selection function (RSF); Discrete choice models (DCM)

## 2.1 Introdução

Inicialmente para melhor entendimento é importante apresentar algumas definições teóricas:

- Os recursos naturais são matérias que a natureza oferece e que permitem às espécies subsistir no planeta. Eles podem ser renováveis (ex.: água, ar, bosques, fauna etc.) ou não renováveis (ex.: petróleo, gás, carvão, espécies silvestres, minerais etc.). As populações animais precisam de uma grande quantidade desses recursos para sua sobrevivência.

- A seleção diferencial dos recursos disponíveis, é um dos principais fatores que permitem a coexistência das espécies, sendo uma prioridade para preservar as espécies em risco de extinção (ROSENZWEIG, 1981).

- Assim, para sustentar as populações animais, são necessárias quantidades adequadas desses recursos. Quando uma espécie seleciona melhor seus recursos, satisfaz com maior qualidade suas exigências vitais com recursos de alta qualidade.

- O uso do recurso é definido como a quantidade deste que é utilizada por um animal em um período de tempo fixo.

- A disponibilidade do recurso é a quantidade acessível ao animal durante aquele mesmo período de tempo.

- Abundância é definida como a quantidade do recurso no ambiente.

Existe uma variação considerável na motivação para dirigir os estudos da Seleção do Recurso (RS). Uma floresta em climax, por exemplo, é vital para a existência da coruja manchada (*Strix occidentalis*) no noroeste dos Estados Unidos, (LAYMON et al., 1985; FORSMAN et al., 1984) ou para a existência do veado de rabo preto (*Odocoileus hemionus*) da Ilha Admiralty, no Alaska (SCHOEN; KIRCHHOFF, 1985).

Outro exemplo da RS se refere à avaliação do impacto das mudanças do habitat. Sob determinadas suposições, a densidade dos animais é proporcional à disponibilidade do recurso no equilíbrio do habitat (FAGEN, 1988). A RS é usada em estudos para identificar os recursos críticos pela população de animais e para predizer a ocorrência das espécies. Freqüentemente, os animais são monitorados individualmente e depois agrupados para estimar os efeitos ao nível da população.

A definição da Função de Seleção do Recurso (RSF), requer que o estudo consista em unidades discretas. A população do recurso disponível é então definida como o conjunto destas unidades disponíveis para o uso do animal ou animais estudados. Por exemplo, se o recurso usado pelo animal é um determinado tipo de inseto, as unidades de recursos são o número de tais insetos disponíveis que podem ser consumidos pelo animal na área onde é examinado o alimento. Neste caso, o estudo da RS poderia estar relacionado, a se os animais, para a sua alimentação, preferem esse tipo de inseto.

O método estatístico mais freqüentemente usado para estimar a RSF é a regressão logística, mas modelos de escolha discreta também têm sido utilizados recentemente. A teoria de modelos de escolha discreta

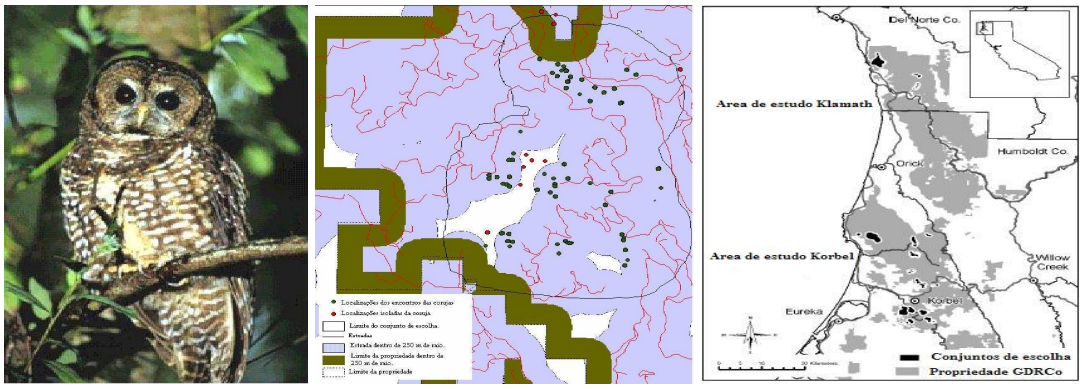
tem sido bem desenvolvida para as análises de dados em escolhas humanas (TRAIN, 2003), mas segundo McDonald et al. (2006), podem ser aplicados também a escolhas do animal com algumas modificações.

## 2.2 Desenvolvimento

Os dados utilizados são provenientes de um estudo sobre a atividade noturna da coruja manchada (*Strix occidentalis*). A região de estudo consistiu em duas áreas discretas (Klamath e Korbel) dentro da propriedade da companhia Green Diamond Resource Company ($GDRCo$), localizada nos condados de Del Norte e Humboldt, no noroeste da California, USA.

Nessa área foram identificados locais de preferência de 28 corujas no período noturno de abril de 1998 a setembro de 2000, usando métodos de rádio-telemetria. Inicialmente, foram capturadas 28 corujas, que receberam um "*chip*" de identificação e foram soltas e monitoradas. Esses "*chips*" permitiram saber a localização por rádio-telemetria (transmissão e processamento de dados a distância) dos animais e com isso identificar suas preferências de habitat.

Dessa forma, foi possível verificar que 5 corujas residiam em Klamath e 23 em Korbel, 46 variáveis explicativas foram igualmente observadas (ver tabela 2.1) resultando em um total de 8.739 observações. Na figura 2.1 observa-se a coruja manchada e o mapa com as localizações dos encontros das corujas (cor verde), assim como as localizações das corujas isoladas, isto é, aquelas que não se movimentaram (cor vermelha) (ver figura 2.1) (RYAN; McDONALD; LAMPHEAR, 2004).




(a) (b) (c)

Figura 2.1 – (a) Coruja manchada (*Strix occidentalis*) e (b-c) ilustração das localizações aleatórias usadas na estimação da Função da Seleção do Habitat (HSF) da espécie nos limites da propriedade Green Diamond

Segundo McDonald et al. (2006), quando os cientistas da fauna aplicam modelos de escolha discreta, eles geralmente supõem que os animais fazem uma série de seleções a partir de conjuntos finitos de unidades de habitat discretas, conhecidas como conjuntos de escolha.

Outras análises da seleção do recurso incluem a regressão logística aplicada para uma amostra que contém as unidades de recurso usadas e não usadas. Supõe-se que as escolhas sejam feitas a partir de um conjunto de unidades de recurso disponíveis.

Modelos de escolha discreta são usualmente aplicados em situações em que *n* conjuntos de unidades de recurso, $N_i, (i = 1, 2, ..., n)$, são definidos como disponíveis para seleção, e de cada um destes conjuntos de escolha é feita a escolha de uma unidade.

O Modelo de Escolha Discreta (DCM ) supõe que a probabilidade de selecionar a *j*-ésima unidade do *i*-ésimo conjunto de escolha é proporcional a,

$$\exp(\boldsymbol{\beta}_1 x_{ij1} + \boldsymbol{\beta}_2 x_{ij2} + \ldots + \boldsymbol{\beta}_p x_{ijp}),$$

em que $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ são coeficientes a serem estimados e $x_{ij1}, \ldots, x_{ijp}$ são os valores de *p* covariáveis medidas na *j*-ésima unidade do *i*-ésimo conjunto de escolha.

Levando em conta estas condições, a probabilidade da *j*-ésima unidade ser selecionada no *i*-ésimo conjunto de escolha é:

$$p_{ij} = \frac{\exp(\boldsymbol{\beta}_1 x_{ij1} + \boldsymbol{\beta}_2 x_{ij2} + \ldots + \boldsymbol{\beta}_p x_{ijp})}{\sum_{k=1}^{n_i} \exp(\boldsymbol{\beta}_1 x_{ik1} + \boldsymbol{\beta}_2 x_{ik2} + \ldots + \boldsymbol{\beta}_p x_{ikp}).} \tag{2.1}$$

Isto é, para S escolhas independentes, a função de verossimilhança é igual à multiplicação das probabilidades de sucesso das escolhas

$$L = l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p) = \prod_{i=1}^{S} (p_{i1})^{y_{i1}} \times (p_{i2})^{y_{i2}} \times \ldots \times (p_{in_i})^{y_{in_i}} \tag{2.2}$$

em que, $y_{ij} = 1$ se o recurso da unidade *j* é escolhido para uso no conjunto de escolha *i* e $y_{ij} = 0$ em caso contrário, $n_i$ é o índice para a unidade escolhida no $i$-ésimo conjunto de escolha, e $p_{ij}$ são os valores dados pela expressão (2.1).

Estimativas de máxima verossimilhança dos parâmetros $\boldsymbol{\beta}$ são obtidas pela maximização de $L$ com respeito a esses parâmetros, fornecendo as estimativas de erros padrão e testes de significância.

Segundo Manly et al., (2002), a Função de Probabilidade da Seleção do Recurso (RSPF ) assume a forma

$$w^*(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_p x_p)}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ldots + \boldsymbol{\beta}_p x_p)}, \tag{2.3}$$

em que $\mathbf{x} = (x_1, \ldots, x_p)$ é o vetor dos valores das variáveis explicativas X. A função logística tem a propriedade desejável de restringir os valores de $w^*(\mathbf{x})$ em 0 ou 1. Caso contrário está função não pode ser utilizada.

Outra justificativa para usar a função logística, em relação a outras abordagens para aproximar a RSPF é o fato dela ser amplamente usada para outras análises estatísticas em biologia, e conseqüentemente, a existência de vários programas de computador atualmente disponíveis para estimar seus parâmetros.

A função estimada,

$$\widehat{w}(\mathbf{x}_{ij}) = \exp(\widehat{\boldsymbol{\beta}_1} x_{ij1} + \widehat{\boldsymbol{\beta}_2} x_{ij2} + \ldots + \widehat{\boldsymbol{\beta}_p} x_{ijp})$$

é então a RSF , dada a probabilidade relativa de uso para diferentes tipos de unidades de recurso. Programas de computador que estimam por máxima verossimilhança os parâmetros de modelos de escolha discreta incluem SAS/Proc PHREG e S-Plus rotina COXPH, (MANLY et al., 2002).

Quando se usa um modelo paramétrico para a probabilidade da RS , os parâmetros são estimados pela máxima verossimilhança. Assim, a quantidade,

$$D \quad = \quad -2\{\log_e(L_M) - \log_e(L_F)\}, \tag{2.4}$$

é chamada deviance, que pode ser usada como uma medida da qualidade do ajuste do modelo. $L_M$ é a máxima verossimilhança para o modelo ajustado, e $L_F(\geq L_M)$ é a verossimilhança para o modelo que se ajusta perfeitamente aos dados. Quando $L_F = L_M$ corresponde ao modelo Nulo (M.N).

Testes Qui-Quadrado em deviance podem ser usados para avaliar, qualquer evidência de probabilidade de uso das parcelas de estudo. Sob certas condições a estatística deviance segue aproximadamente a distribuição Qui-Quadrado com os graus de liberdade (df) definidos pelo número de observações menos o número de parâmetros estimados que o modelo atual é correto. A deviance é análoga à soma de quadrados para a regressão dos modelos de análise de variância.

Uma importante medida de qualidade do ajuste do modelo é o Critério de Informação Akaike (AIC). O AIC para o modelo é definido como,

$$AIC \quad = \quad -2 \quad \{\log_e(L_M)\} + 2p, \tag{2.5}$$

em que *p* é o número de parâmetros desconhecidos no modelo que deve ser estimado (AKAIKE, 1974).

Inicialmente, localizamos em um gráfico as variables "$X\_COORD$" e "$Y\_COORD$", que são as localizações de todas as corujas manchadas, como observa-se na figura (2.2). Como elas não freqüentaram todos os lugares da região de estudo, e cada coruja manchada fixa-se num território, isso pode evidenciar o fato delas não usarem todas as variáveis para a análise.

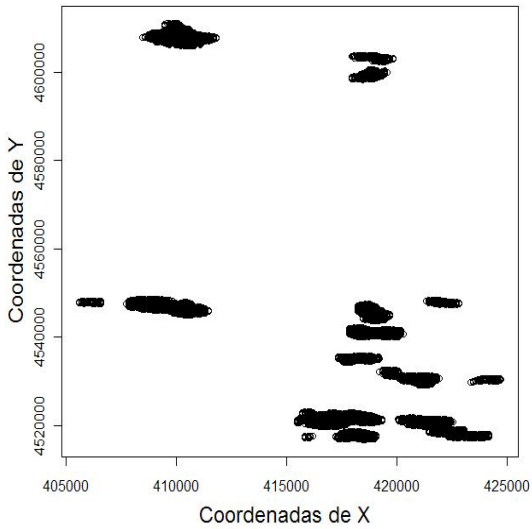



Figura 2.2 – Localizações em coordenadas X e Y da Coruja Manchada

Tabela 2.1 – Variáveis explanatórias do habitat da coruja manchada (*Strix occidentalis*)

| Variável | Descrição |
|---|---|
| ID | Número índice da observação |
| USED | Indicador para o ponto usado $(1)$, ou ponto disponível $(0)$ |
| Xcoord | UTM$^{(1)}$ |
| Y coord | UTM$^{(1)}$ |
| OWL | Identificação única da coruja |
| SPECENT2 | Variável indicadora para espécies dominantes Redwood. |
| SPECENT3 | Variável indicadora para espécies dominantes Whitewood. |
| aclc2 | Variável indicadora para a classe de idade da parcela entre 6 e 20 anos. |
| aclc3 | Variável indicadora para a classe de idade da parcela entre 21 e 40 anos. |
| aclc4 | Variável indicadora para a classe de idade da parcela de 41 anos ou mais. |
| acln2 | Variável indicadora da classe de idade do bosque ao redor do indivíduo entre 6 e 20 anos. |
| acln3 | Variável indicadora da classe de idade do bosque ao redor do indivíduo entre 21 e 40 anos. |
| acln4 | Variável indicadora da classe de idade do bosque ao redor do indivíduo entre 41 anos ou mais. |
| (aclc_×acln_) | 9 Variáveis indicadoras da interação classe de idade da parcela e do bosque ao redor do indivíduo |
| acl_p2 | Proporção da zona tampão de bosque de idade entre 6 e 20 anos. |
| acl_p3 | Proporção da zona tampão de bosque de idade entre 21 e 40 anos. |
| acl_p4 | Proporção da zona tampão de bosque de idade entre 41 anos ou mais. |
| age_cent | Idade de bosque em anos. |
| acl_dne | Distância da borda mais próxima (em pés)$^{(2)}$. |
| acl_te | Perímetro total da borda (em pés). |
| acl_ed | Densidade da borda da zona tampão (em pés), aclte/área da zona tampão (em pés). |
| acl_mps | Tamanho médio da mancha da zona tampão (em pés). |
| acl_np | Número de manchas dentro da zona tampão. |
| acl_pd | Densidade da mancha na zona tampão (n/100 acres$^{(3)}$), aclnp área da zona tampão. |
| acl_pscv | Porcentagem do coeficiente de variação do tamanho da mancha. |
| drd_cent | Distância à estrada principal (em pés). |
| hgt_cent | Altura média das árvores no bosque (em pés). |
| phw_cent | Porcentagem de (Hardwood) (madeira dura) |
| prs_cent | Porcentagem do remanescente após o corte das árvores |
| prw_cent | Porcentagem de Redwood. |
| pww_cent | Porcentagem de Whitewood. |
| rwd_cent | Área basal Redwood ( pés/acre) |
| wwb_cent | Área basal Whitewood (pés/acre) |
| hwd_cent | Área basal Hardwood (pés/acre) |
| res_cent | Área basal de remanescente depois de cortar as árvores (pés/acre) |
| tba_cent | Total da área basal (pés/acre). |
| slp_cent | Porcentagem de inclinação do terreno. |

Fonte: Western EcoSystems Technology - Cheyenne - USA

$^{(1)}$ UTM é o sistema de posicionamento global usado para obter latitude, longitude, e altitude dos dados em qualquer parte do mundo (BURROUGH, 2004).

$^{(2)}$ Um pé equivale a 30,48cm.

$^{(3)}$ Um acre equivale a 4047 $m^2$.

O tipo de delineamento usado no conjunto de dados da coruja manchada foi o delineamento tipo II, em que, os animais são identificados individualmente e o uso das unidades do recurso é medido para cada um, mas a disponibilidade do recurso é medida para o total da população. O protocolo de amostragem foi o $C$, em que, as unidades de recurso usadas e não usadas são amostradas independentemente, (MANLY et al., 2002).

Para fazer as comparações da regressão logística e o modelo de escolha discreta foi selecionada uma amostra aleatória de 390 observações, dos dados da coruja manchada com uma escolha. Para a seleção das variáveis usou-se o AIC. Para obter as estimativas da regressão logística foram utilizados os sistemas computacionais MINITAB (1997), e R (2006). Já para a estimação dos parâmetros do DCM usou-se a linguagem de programação Fortran (FORTRAN 77, 1995).

Tabela 2.2 – Estimativas dos parâmetros de Regressão Logística e do Modelo de escolha discreta (DCM), para a seleção do habitat da coruja manchada

| | Regressão logística | | | Modelo de escolha discreta | |
|---|---|---|---|---|---|
| Variáveis | Coeficientes | Erro-padrão | Valor-p | Coeficientes | Erro-padrão |
| Constante | -3,9758 | 1,2618 | 0,002 | | |
| aclc2 | -0,3606 | 1,1461 | 0,753 | -0,0747 | 1,2672 |
| aclc3 | -1,0449 | 1,4836 | 0,481 | -1,1875 | 1,5949 |
| aclc4 | 0,4724 | 1,1233 | 0,674 | 0,7734 | 1,2807 |
| acl_ed | 0,0127 | 0,0058 | 0,029 | 0,0176 | 0,0077 |
| hgt_cent | 0,0104 | 0,0049 | 0,034 | 0,0147 | 0,0064 |
| slp_cent | -0,0171 | 0,0073 | 0,019 | -0,0289 | 0,0103 |

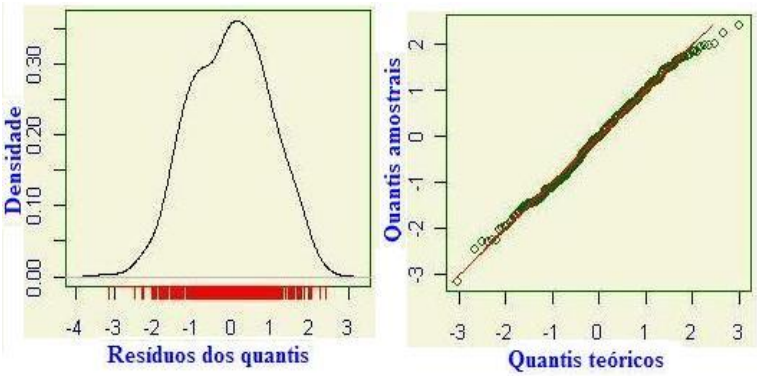



Figura 2.3 – Distribuição de freqüências dos resíduos e Q-Q plot para a distribuição Binomial com a função de ligação logit

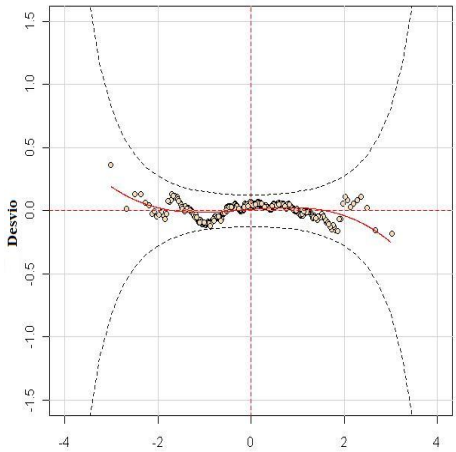



Figura 2.4 – Gráfico “worm” da distribuição Binomial com a função de ligação logit

Na figura (2.3), observa-se o ajuste da distribuição binomial com função de ligação logit para as variáveis selecionadas. Também, na figura (2.4), o gráfico “worm” é uma ferramenta de diagnóstico geral para as análises dos resíduos, o eixo vertical são as diferenças entre as distribuições teórica e empírica. O gráfico “worm” deve apresentar a forma de uma corda, indicando, em nosso caso, que os dados têm a distribuição binomial, e podemos observar os pontos consecutivos (BUUREN; FREDRICKS, 2001).

Tabela 2.3 – Comparação dos modelos pelo teste Qui-Quadrado

| | Regressão Logística (RL) | Modelo de escolha discreta (DCM ) |
|---|---|---|
| Deviance | 163,78 (383gl) | 107,02 (384 gl) |
| AIC | 177,78 | 119,02 |

Na tabela (2.2), são apresentadas as estimativas dos parâmetros da regressão logística e do modelo de escolha discreta. Na tabela (2.3), vê-se que a deviance do DCM reduz -56,76 em relação a deviance da RL, e o critério AIC do DCM reduz com respeito a RL em -58,76. Nota-se que a regressão logística possui um grau de liberdade a menos em relação ao DCM , devido a que este último não possui intercepto.

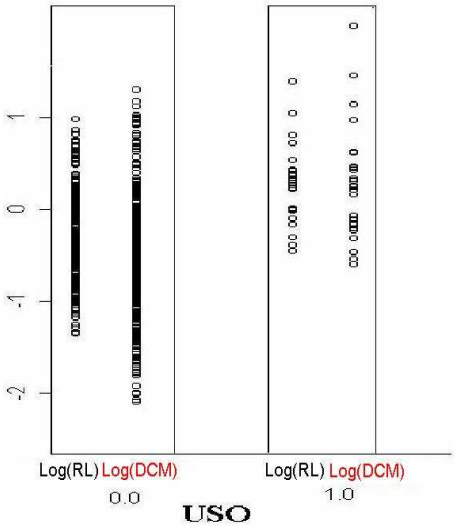



Figura 2.5 – Comparação dos usos da Coruja manchada (*Strix occidentalis*) com a regressão logística (RL) e com o modelo de escolha discreta (DCM )

Na figura (2.5), observa-se no eixo *X*, se a coruja não freqüentou o lugar (0) e se ela usou (1), nos comparamos as estimativas do logaritmo da RL e o logaritmo do DCM no eixo *Y*, mas os nomes foram ubicadas no eixo *X* para melhor apresentação da figura, e podemos dizer que a coruja freqüentou muitos lugares, embora tenha usado poucos deles. A situação é mais complicada para amostras aleatórias independentes que são tomadas separadamente em diferentes tipos de unidades: disponíveis, usadas e não usadas. Neste caso a Regressão logística pode ainda ser usada, no entanto, precisa de especial justificativa dependendo dos tipos de amostras envolvidas. Em nosso caso, para amostras independentes de unidades usadas e disponíveis, supondo a população de unidades disponíveis de tamanho *N*, com a i-ésima unidade assumindo os valores $\mathbf{x}_i$=$(x_{i1}, \ldots, x_{ip})$, para as variáveis $X_1$ a $Xp$ e a probabilidade relativa de uso para diferentes unidades de

recurso é correspondente a:

$$w^*(\mathbf{x}_i) = \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_p x_{ip}) \tag{2.6}$$

O plano de amostragem é tal que cada unidade disponível tenha uma probabilidade $P_a$ de ser amostrada, e cada unidade usada tenha uma probabilidade $P_{\bar{u}}$ de ser amostrada, com a amostra de unidades disponíveis selecionada primeiro, sem reposição, de modo que as unidades nessa amostra não possam aparecer na amostra de unidades não usadas.

Neste caso, a probabilidade de uma unidade ser usada e amostrada é $(1-P_a)w^*(\mathbf{x}_i)P_{\bar{u}}$ e a probabilidade de uma unidade estar na amostra de unidades usadas ou na amostra de unidades disponíveis é dada por:

$$Prob(\textbf{i}\text{-ésima } \textbf{u}\text{nidade } \textbf{a}\text{mostrada}) \;=\; P_a + (1-P_a)w^*(\mathbf{x}_i)P_{\bar{u}}.$$

Assim, a probabilidade da $i$-ésima unidade estar na amostra de unidades usadas, dado que ela foi amostrada é dada por:

$$Prob(\text{i-ésima unidade usada} \mid \text{amostrada}) = Prob(\text{usada e amostrada}) \;/\; Prob(\text{amostrada})$$

$$= \frac{(1-P_a)w^*(\mathbf{x}_i)P_{\bar{u}}}{P_a + (1-P_a)w^*(\mathbf{x}_i)P_{\bar{u}}}. \tag{2.7}$$

Dado que a RSPF definida na expressão (2.6) assume uma forma exponencial particular, a probabilidade da expressão (2.7) pode ser escrita também como:

$$\boldsymbol{\tau}(\mathbf{x}_i) = \frac{\exp\left\{\log\left[\frac{\left(1-P_a\right)P_{\bar{u}}}{P_a}\right] + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_p x_{ip}\right\}}{1 + \exp\left\{\log\left[\frac{\left(1-P_a\right)P_{\bar{u}}}{P_a}\right] + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_p x_{ip}\right\}} \tag{2.8}$$

que corresponde a uma expressão da regressão logística na qual, o parâmetro $\boldsymbol{\beta}_0$ é modificado para,

$$\boldsymbol{\beta}_0^{'} = \log\left[\frac{(1-P_a)P_{\bar{u}}}{P_a}\right] + \boldsymbol{\beta}_0 \tag{2.9}$$

para levar em consideração as probabilidades de amostragem das unidades usadas e disponíveis. Assumindo independência das observações, $\boldsymbol{\tau}(\mathbf{x}_i)$ representa a probabilidade de observar a unidade de recurso *i*, como sendo usada, e a probabilidade de observar tal unidade, como sendo disponível é então dada por $1 - \boldsymbol{\tau}(\mathbf{x}_i)$.

Seja $y_i$ o indicador do uso ou não uso de uma unidade amostrada, tal que $y_i$ = 0, se a unidade amostrada *i* está na amostra de unidades disponíveis e $y_i$=1, se a unidade *i* está na amostra de unidades usadas. A probabilidade de observar a unidade *i* pode então ser escrita como,

$$L_i = \boldsymbol{\tau}(\mathbf{x}_i)^{y_i}(1 - \boldsymbol{\tau}(\mathbf{x}_i))^{1-y_i} \tag{2.10}$$

e o logaritmo da verossimilhança de observar a amostra completa é:

$$\log\{L(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)\} = \sum_{i=1}^{n} \log L_i \tag{2.11}$$

$$\log\{L(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)\} = \sum_{i=1}^{n} y_i \log\{\boldsymbol{\tau}(\mathbf{x}_i)\} + (1 - y_i)\log\{1 - \boldsymbol{\tau}(\mathbf{x}_i)\} \tag{2.12}$$

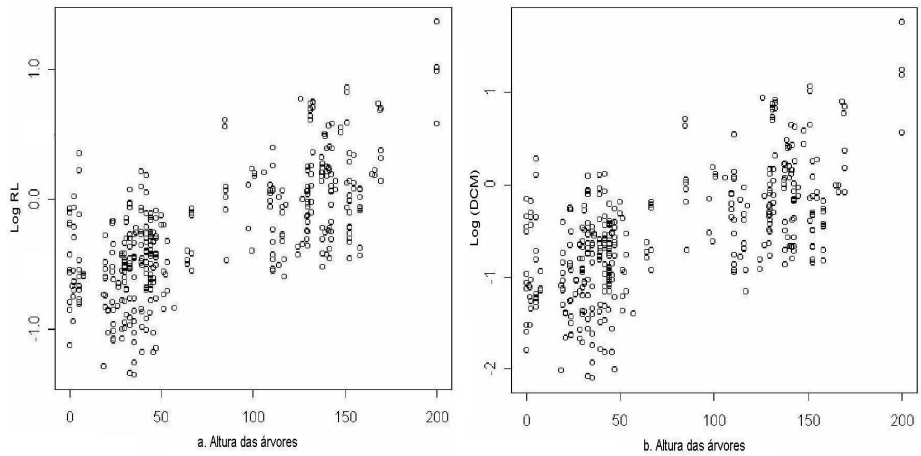
Programas de computador de regressão logística podem ser usados para estimar os coeficientes $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ da função logaritmo linear da expressão (2.6).

O fato da constante da regressão logística $\boldsymbol{\beta}_0^{'}$ assumir a forma da expressão (2.9), significa que se as probabilidades de amostragem $P_u$ e $P_a$ são conhecidas, então o parâmetro $\boldsymbol{\beta}_0$ da RSPF na expressão (2.6) pode ser estimado subtraindo a quantidade $\log[(1 - P_a)P_u/P_a]$ da constante estimada $\boldsymbol{\beta}_0^{'}$ na equação da regressão logística. Se as frações de amostragem não são conhecidas, então $\boldsymbol{\beta}_0$ não pode ser estimado, mas ainda é possível estimar a RSF ,

$$w^*(\mathbf{x}) = \exp(\boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) \tag{2.13}$$

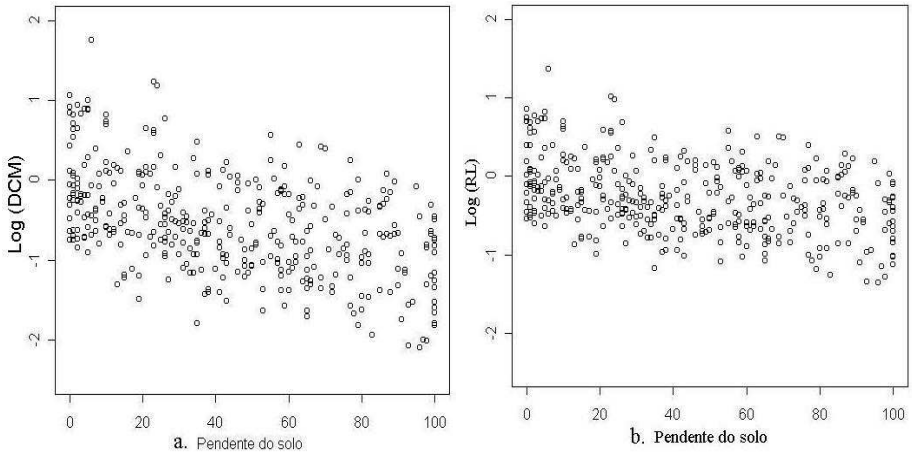
e usar esta função para comparar as unidades de recurso.

Nota-se que as probabilidades corretas de uso ou as probabilidades relativas de uso, são obtidas substituindo as estimativas de $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$ nas funções logaritmo lineares das expressões (2.6) ou (2.13). As probabilidades obtidas usando programas de computador para ajustar a regressão logística, $\boldsymbol{\tau}(\mathbf{x}_i)$ na expressão (2.8) não são estimativas corretas para seleção da probabilidade de recurso, $w^*(\mathbf{x}_i)$, ou para a função da seleção do recurso $w(\mathbf{x}_i)$, devido a que as outras estimativas levam em conta o intercepto, sendo este parâmetro não muito útil, em situações de ecologia. Observa-se na figura (2.6), a similaridade nos gráficos de regressão logística e DCM. Na figura (2.7), da pendente do solo observa-se um padrão diferente no gráfico do modelo de escolha discreta.




(a) (b)

Figura 2.6 – a. Altura das árvores *vs* Log (Regressão Logística (RL)), b. Altura das árvores *vs* Log (Modelo de escolha discreta (DCM ))




(a) (b)

Figura 2.7 – a. Pendente do solo *vs* Log (Modelo de escolha discreta (DCM )), b. Pendente do solo *vs* Log (Regressão logística (RL))

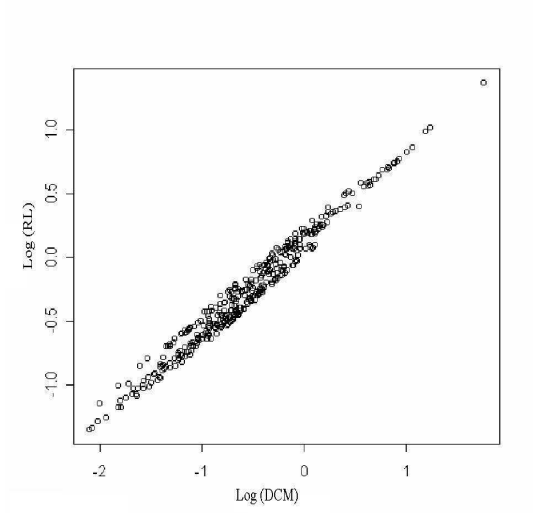



Figura 2.8 – Log(Modelo de Escolha Discreta (DCM )) *vs* Log(Regressão Logística (RL))

Ao fazer as estimativas de cada uma das escolhas da coruja observamos na tabela 2.2 que, os coeficientes estimados pela regressão logística diferem com respeito aos coeficientes de todas as escolhas da coruja. O mesmo acontece com o DCM nas estimativas dos parâmetros do modelo de escolha, com respeito a todas as corujas. Observando-se na figura (2.8) o melhor ajuste apresentado pelo logaritmo (DCM), com uma escolha (VERGARA et al., (2006, 2006a)).

## 2.3 Conclusões

i ) A RSF foi estimada com sucesso pela regressão logística e é usada em estudos para identificar os recursos críticos pela população de animais e predizer a ocorrência das espécies.

ii ) Observa-se que ao fazer o ajuste na regressão logística na base de dados da coruja manchada e o Modelo de Escolha Discreta (DCM ), apresenta uma melhor alternativa para uma escolha da coruja.

iii ) As estimativas dos parâmetros na regressão logística e no DCM com uma escolha, apresentam boa similaridade.

iv ) As análises feitas de todas as escolhas individualmente diferem das análises feitas escolha por escolha, o que justifica o uso de modelos de efeitos aleatórios para todos os animais considerados simultaneamente. Portanto, a regressão logística e os DCM podem ser generalizados para incluir efeitos aleatórios.

**Referências**

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, New York, v. 19, p. 716- 723, 1974.

BUUREN, S. V; FREDRICKS, M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistics in Medicine**, Chichester, v. 20, p. 1259-1277, 2001.

BURROUGH, P. **Principles of geographical information systems**. Oxford: New York : Oxford University Press, 2004. 333 p.

FAGEN, R. Population effects of habitat change: a quantitative assessment. **Journal of Wildlife Management**, Cheyenne, v. 52, p. 41-46, 1988.

FORSMAN, E.D.; MESLOW, E.; WIGHT, H.M. Distribution and biology of the spotted owl in Oregon. **Wildlife Monographs**, Bethesda, v. 87, p. 1-64, 1984.

FORTRAN 77. **Programmer's Guide.** California, 1995. 1 CD-ROM.

LAYMON, S. A.; SALWASSER, H.; BARRET, R. H. **Habitat suitability index models**: spotted owl. United States Fish and Wildlife Service, Washington, D.C: Biological Services Program, 1985. (Report 82(10.113)).

MANLY, B.F.J.; McDONALD, L.L.; THOMAS, D.L.; McDONALD, T.L.; ERICKSON, W.P. **Resource selection by animals.** 2nd ed. London: Kluwer Academic Publishers, 2002. 221 p.

McDONALD, T.L.; MANLY, B.F.J.; NIELSON, R.M.; DILLER, L.V. Discrete choice models with a spotted owl example. **Journal of Wildlife Management**, Cheyenne, v. 70, p. 375-83, 2006.

MINITAB. **Minitab user's guide 2**: data analysis and quality tools. Pennsylvania, 1997.

**R**: a language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0, Vienna, Austria, 2006. Disponível em: <http://www.R-project.org>. Acesso em: 01 maio. 2006.

ROSENZWEIG, M.L. A theory of habitat selection. **Ecology**, Brooklyn, v. 62, p. 327-335, 1981.

RYAN, N.; McDONALD, T.L; LAMPHEAR, D. **Northern spotted owl nighttime site selection Model**. Cheyenne: s.e.d. 2004. 38 p. (Report Western EcoSystems Technology)

SCHOEN, J. W; KIRSCHHOFF, M. D. Seasonal distribution and home range patterns of Sitka black-tailed deer on Admiralty Island, southeast Alaska. **Journal of Wildlife Management**, Cheyenne, v. 49, p. 96-103, 1985.

TRAIN, K. **Discrete choice methods with simulation.** Cambridge: Cambridge University PRESS, 2003. 329 p.

VERGARA, C.S.; DIAS, C.T.S.; MANLY, B.F.J. Regressão Logística e Modelos de Escolha Discreta na Seleção do Habitat pela Coruja Manchada (Strix occidentalis). In: 51º REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA (RBRAS) E 12ª SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA (SEAGRO)., 2006, Botucatu. **Resumos**... Botucatu: UNESP, 2006. Resumo 120.

VERGARA, C.S.; DIAS, C.T.S.; MANLY, B.F.J. Modelos de Regressão Logística e de Escolha Discreta na Seleção do Habitat por animais. In: 17º SIMPOSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA (SINAPE), 2006, Caxambu. **Comunicação oral**... Caxambu: ABE, 2006a.

# 3 MÉTODOS DE BOOTSTRAP PARA A ESTIMAÇÃO DE VIÉS E VARIÂNCIAS NA FUNÇÃO DE PROBABILIDADE DE SELEÇÃO DO RECURSO (RSPF )


**Resumo**

Funções de Seleção do Recurso RSFs (Resource Selection Functions), são geralmente usadas para quantificar quanto os animais são seletivos durante o período de uso do habitat ou alimento. De início a Função da Probabilidade da Seleção de Recurso RSPF (Resource Selection Probability Function) pode ser claramente estimada se $N$, o número total da unidade de recurso na população, e $n_1$, o número total de unidades usadas durante o período de estudo, são ambas conhecidos e pequenos. A aproximação da RSPF pode ser então estimada usando qualquer programa de regressão logística padrão, mas neste caso, as variâncias das estimativas dos parâmetros serão muito pequenas e a qualidade da estatística ajustada não será boa. Para superar estes problemas, a reamostragem bootstrap dos dados é então aqui proposta. Também é feita uma discussão sobre algumas potenciais limitações do uso da regressão logística para estimar as RSPFs . O método para estimar a RSPF descrito aqui, tem aplicações potenciais em medicina, ecologia e outras áreas.


Palavras-chave: Funções da seleção do recurso (RSFs); Função da probabilidade da seleção de recurso (RSPF ); Reamostragem bootstrap; Regressão logística


**Abstract**

Resource selection functions (RSFs) are used for quantify how animals are selective in the use of the habitat period or food. A Resource Selection Probability Function (RSPF) can be estimated if *N*, the total number of units in the population, and $n_1$ the total number of used units in the study period are both known and small. An approximation of the RSPF can then be estimated using any standard program for logistic regression but the variances of the estimates of the parameters are too small. Three methods of bootstrap sampling, parametric, non-parametric and a modified parametric method are proposed for the estimation of variances, with a discussion about the limitations of logistic regression for estimating RSPF. The method for estimating the RSPF described here has potential applications in medicine, ecology and other areas.


Keywords: Resource selection functions (RSFs); Resource selection probability function (RSPF); Bootstrap; Logistic regression

## 3.1 Introdução

Funções de Seleção do Recurso (Resource selection functions RSFs ), são geralmente usadas para quantificar o quanto os animais são seletivos durante o período de uso do habitat ou alimento, como discutido por Strickland; McDonald (2006), e outros autores citados nesta publicação. Estas funções foram desenvolvidas por volta de 1980 como uma generalização das funções ajustadas, que eram geralmente usadas para o estudo da seleção natural sobre as populações de animais (MANLY, 1985). Funções ajustadas estão relacionadas com a probabilidade de seleção da sobrevivência apesar das RSFs estarem relacionadas

com a probabilidade dos recursos serem usados (McDONALD et al., 1990). O uso da RSF depende, em uma população de *N* unidades de recurso a serem definidos, em que cada unidade é definida pelo valor de *p* variáveis $X_1$ a $X_p$.

As unidades de recurso podem ser a quantidade de alimento, em cada caso, dentro das quais as variáveis *X* podem ser descritas com base em seu tamanho e cor. Alternativamente, as unidades de recurso podem ser unidades do habitat como uma parcela de um hectare dentro de um grande parque nacional, neste caso as variáveis *X* poderiam descrever a distância da parcela de terra à água, a elevação, e a vegetação dominante na parcela. Para o restante deste capítulo, somente a seleção de habitat será discutida, mas em geral o que foi dito também se aplica para a seleção dos alimentos.

O número de unidades de recurso *N*, pode ser muito grande ou até infinito e, pode também não ser conhecido. Por exemplo, Manly et al. (2002), descrevem a situação de 24 grupos de vegetação que foram selecionados por conter ninhos de pardais que durante um período de dois anos, foram comparados com 25 grupos de vegetação selecionados aleatoriamente na área de estudo. Neste caso, as unidades de recurso são os grupos de vegetação, dentro da área de estudo, sendo o número total delas desconhecido, mas provavelmente muito grande.

Em alguns estudos as unidades de recurso podem ser consideradas como sendo pontos dentro da área de estudo, onde os animais são encontrados. Neste caso poderia haver um número infinito de unidades de recurso dentro da população. Isto pode de qualquer forma ser evitado, se considerarmos a área de estudo consistindo em *N* parcelas de terra não sobrepostas, com a parcela que está sendo usada, se um ou mais animais são encontrados dentro da parcela. Neste caso o valor de *N* poderia depender do tamanho da parcela. Por exemplo, se as parcelas são de um quilômetro quadrado, então pode haver milhares destas parcelas na área de estudo, enquanto que se as parcelas são de um metro quadrado, *N* poderia ser alguns milhões de vezes maior.

Qualquer que seja a definição da unidade do recurso, a RSF é definida como qualquer função $w(x_1, \ldots, \ x_p)$, em que esta é proporcional à probabilidade desta unidade com valores de $x_1$ a $x_p$ para as variáveis $X_1$ a $X_p$ usadas durante a execução do estudo. Enquanto que a função da probabilidade da seleção do recurso (RSPF ) é definida como a função $w^*(x_1, \ldots, \ x_p)$ dando a probabilidade atual de uso para a unidade com valores de $x_1$ a $x_p$ para as variáveis de $X_1$ a $X_p$.

De qualquer maneira, em situações semelhantes às do exemplo anterior dos pardais, onde o número total de grupos de vegetação usados em dois anos do período de estudo, e o número total de grupos de vegetação na a área de estudo são ambos desconhecidos, parece ser irreal esperar uma estimativa com a probabilidade real de uso. Em casos semelhantes, esta RSF poderia ainda ser estimada usando a amostra de unidades usadas. A amostra de unidades disponíveis e esta função seria tudo o que é necessário para quantificar o processo de seleção (MANLY et al., 2002).

Em princípio, a RSPF pode ser claramente estimada se *N*, o número total de unidades de recurso na população, e $n_1$, o número total de unidades usadas durante o período de estudo, são ambas conhecidas. Em muitos estudos não é difícil determinar *N*. Por exemplo, se a unidade de recurso a ser considerada é uma parcela de um hectare dentro de um parque nacional então, é fácil calcular o número total de parcelas dentro do parque. De outro lado, determinar $n_1$ freqüentemente não é tão fácil. Por exemplo, no estudo da localização dos pássaros por freqüência de rádio, o movimento deles poderia ser registrado uma vez por dia.

Então, isso fornece o número de parcelas observadas a serem usadas, mas não têm um meio de conhecer a posição onde o pássaro ficou quando não foi registrado. Em vista disso, propomos que este problema seja estudado pela definição da unidade de recurso usada naquelas que são registradas, como sendo usadas quando as observações de uso são feitas. Isso então fixa o valor de $n_1$ e não há uma resposta sobre o que isso significa para a unidade a ser usada. Isso também significa que todas as $N - n_1$ unidades, dentro da população de unidades, não são observadas para o uso pois, pela definição, são as unidades não usadas.

Dada esta situação, é possível no início estimar a RSPF usando métodos padrões. Assumimos, de qualquer maneira, que para *N* grande, não seja possível adicionar os dados sobre todas as $N - n_1$ unidades não usadas em programa de computador. Por exemplo, poderiam ser vários milhões destas unidades não usadas. Na próxima seção descrevemos um método simples para estimar a RSPF baseada em $n_1$ unidades observadas a serem mais usadas, em grandes amostras de $n_2$ unidades não usadas, retiradas, por exemplo, do sistema de informação geográfica (GIS). Regressão logística é atualmente usada, mas a aproximação aqui proposta pode ser usada igualmente com outra função, descrevendo a relação entre a probabilidade de uso e as variáveis $X_1$ a $X_p$ medidas na unidade de recurso.

A aproximação da RSPF pode ser estimada usando qualquer programa de regressão logística padrão, mas as variâncias das estimativas dos parâmetros, que são as saídas dos programas, serão também pequenas e a qualidade da estatística ajustada não seria boa. Também, existe algum viés devido ao tamanho amostral pequeno usado na estimativa do parâmetro. Para superar estes problemas, a reamostragem bootstrap de dados é então proposta. Três destes métodos são descritos e estudos de simulação indicam que um destes métodos fornece resultados satisfatórios.

Conclui-se, com a discussão sobre algumas limitações potenciais do uso da regressão logística para estimar a RSPF que por exemplo, se a probabilidade de uso para unidades em um ano de estudo é dada exatamente pela expressão da regressão logística e, esta função também contém um segundo ano de estudo, então a função RSPF descreve a probabilidade de uso para dois anos e, portanto, não pode ser a regressão logística, sendo ela seria correta só para um tempo de duração.

## 3.2 Desenvolvimento

### 3.2.1 Método para a estimar a RSPF

O modelo de regressão logística é uma generalização da regressão linear múltipla, sendo usado freqüentemente nos estudos de seleção de recursos. Por exemplo, em uma situação simples, podem-se apresentar dois tipos de alimentos, *A* e *B*, ao animal em *n* ocasiões diferentes. A escolha do alimento *A*, pode ser considerada como sucesso e então observa-se o número de sucessos em *n* tentativas. Neste procedimento, repetido em *m* situações sob diferentes condições ambientais (alimentos disponíveis, condições de luz etc.), é de interesse conhecer como a probabilidade de escolher *A* está relacionada às várias condições ambientais. Com essas condições experimentais, o modelo de probabilidade de sucessos é dado pela expressão:

$$\pi = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}$$

em que $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$,..., $\boldsymbol{\beta}_p$ são as constantes a serem estimadas a partir dos dados disponíveis e $X_1$,..., $X_p$ são as variáveis explicativas que expressam as condições ambientais. Usualmente, supõe-se que o número de sucessos observados em *n* ensaios segue uma distribuição binomial com média $n\pi$ e variância $n\pi(1-\pi)$, o que implica que o resultado de um ensaio é independente do resultado de qualquer outro ensaio. Para ajustar o modelo de regressão logística aos dados, usualmente utiliza-se o número de sucessos observados sob *m* diferentes condições, e os valores das variáveis explicativas $X_1$,..., $X_p$, com um conjunto de valores para cada uma das *m* condições. O método da máxima verossimilhança é usado para obter as estimativas dos valores de $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_p$.

Considerando a situação da regressão logística padrão, suponha uma amostra aleatória de unidades de recurso em que cada unidade é classificada como usada ou não usada. Então temos $n_1$ unidades usadas e $n_2$ unidades não usadas, e a probabilidade de uso para uma unidade é modelada pela função logística,

$$P(X_1 = x_1, \ldots,\ X_p = x_p) = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)} \tag{3.1}$$

A regressão logística pode ser usada para relacionar a probabilidade de uso de *p* variáveis explicativas $x_1,\ \ldots,\ x_p$, medidas na unidade de recurso. A RSPF é como a expressão (3.1)

$$w^*(\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}$$

em que $\mathbf{x} = (x_1, \ldots, x_p)$ é o vetor dos valores das variáveis explicativas. A função logística tem a propriedade desejável de restringir valores de $w^*(\mathbf{x})$ entre 0 e 1; caso contrário não se utiliza tal função (MANLY et al., 2002).

A probabilidade de fracasso da RSPF é,

$$1 - w^*(\mathbf{x}) = \frac{1}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)} \tag{3.2}$$

A RSPF para a dependência de $w^*(\mathbf{x})$ em valores de *p* variáveis explicativas, $x_1,\ \ldots,\ x_p$, associada com esta observação do exemplo é:

$$logit(w^*(\mathbf{x})) = \log\left(\frac{w^*(\mathbf{x})}{1 - w^*(\mathbf{x})}\right) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p \tag{3.3}$$

e

$$\widehat{w}^*(\mathbf{x}) = \frac{\exp(\widehat{\boldsymbol{\beta}}_0 + \widehat{\boldsymbol{\beta}}_1 x_1 + \ldots + \widehat{\boldsymbol{\beta}}_p x_p)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p)}. \tag{3.4}$$

### 3.2.2 Uma nova estimação de máxima verossimilhança

O método de estimação da RSPF é usualmente o de máxima verossimilhança. Assume-se que as unidades de 1 a $n_1$ são usadas e as unidades $n_1 + 1$ a $n_1 + n_2$ não são usadas. Assim, a função de verossimilhança da probabilidade dos dados observados, dada pela expressão (3.1), é apresentada, assim,

$$L = \Big\{\prod_{i=1}^{n_1} P_i\Big\}\Big\{\prod_{i=n_1+1}^{n_1+n_2} (1 - P_i)\Big\}, \tag{3.5}$$

em que $P_i$ é a probabilidade de uso da função logística para a unidade de recurso *i*, como definida na expressão (3.1).

Usualmente o que é maximizado é o logaritmo natural da função de verossimilhança, isto é,

$$\log(L) = \sum_{i=1}^{n_1} \log(P_i) + \sum_{i=n_1+1}^{n_1+n_2} \log(1 - P_i). \tag{3.6}$$

$y_i$, corresponde as unidades, 0, se não foram usadas, e 1 se são usadas. A derivada do logaritmo da função de verossimilhança em relação aos *p+1* parâmetros $\boldsymbol{\beta}$ desconhecidos é:

$$l = \prod_{i=1}^{n_1+n_2} \left(P_i\right)^{y_i} \left(1 - P_i\right)^{(1-y_i)}$$

$$\ell = \log(l) = \sum_{i=1}^{n_1+n_2} \left[ y_i \log(P_i) + (1 - y_i) \log(1 - P_i) \right]$$

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n_1+n_2} \left[ \left( \frac{y_i}{P_i} \right) \frac{dP_i}{d\boldsymbol{\beta}_j} - \left( \frac{1 - y_i}{1 - P_i} \right) \frac{dP_i}{d\boldsymbol{\beta}_j} \right] \quad i = 1, \ldots, n_1 + n_2, \quad j = 1, \ldots, P + 1 \\ &= \sum_{i=1}^{n_1+n_2} \left[ \frac{y_i - y_i P_i - P_i + y_i P_i}{P_i(1 - P_i)} \right] \frac{dP_i}{d\boldsymbol{\beta}_j} \end{aligned}$$

$$= \sum_{i=1}^{n_1+n_2} \left[ \frac{y_i - P_i}{P_i(1 - P_i)} \right] \frac{dP_i}{d\beta_j} \tag{3.7}$$

$$\begin{aligned} &\approx \sum_{i=1}^{n_1} \left[ \frac{y_i}{P_i(1 - P_i)} \right] \frac{dP_i}{d\boldsymbol{\beta}_j} \\ &- \frac{N}{m} \sum_{i=n_1+1}^{n_1+n_2} \left[ \frac{P_i}{P_i(1 - P_i)} \right] \frac{dP_i}{d\boldsymbol{\beta}_j}. \end{aligned}$$

O primeiro somatório do lado direito da aproximação é relativo às unidades usadas e o segundo somatório é relativo às amostras desconhecidas. Em grandes amostras, esta diferença poderia ser usualmente pequena, mas a diferença pode ser grande quando o número de observações no conjunto de dados é pequeno. Usando a expressão (3.1) para as derivadas com respeito a $\boldsymbol{\beta_j}$, temos que,

$$\frac{\partial P_i}{\partial \boldsymbol{\beta}_j} = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) x_i}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)} - \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) x_i}{[1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)]^2}$$

$$\frac{\partial P_i}{\partial \boldsymbol{\beta}_j} = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) x_i (1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p))}{(1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p))^2}$$

$$- \quad \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) x_i}{[1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)]^2}$$

$$\frac{\partial P_i}{\partial \boldsymbol{\beta}_j} = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) x_i}{[1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)]^2}$$

$$\frac{\partial P_i}{\partial \boldsymbol{\beta}_j} = P_i[1 - P_i]x_i, \quad \text{em} \quad \text{que} \;\; x_0 = 1.$$

A derivada da função do logaritmo da verossimilhança com relação aos $p + 1$ parâmetros $\boldsymbol{\beta}$ desconhecidos é dada pela expressão (3.7) que ao substituir a expressão anterior fica:

$$\begin{aligned} \frac{\partial \ell}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n_1+n_2} \frac{y_i - P_i}{P_i(1 - P_i)} [P_i(1 - P_i)] x_i \\ \frac{\partial \ell}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n_1+n_2} [y_i - P_i] x_i \end{aligned}$$

se, $y_i = 1$, unidades usadas, a derivada fica:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}_j} = \sum_{i=1}^{n_1+n_2} [1 - P_i] x_i.$$

Avaliando estas derivadas de $\hat{\boldsymbol{\beta}}$ e igualando a zero, o conjunto das *k+1* equações não lineares de parâmetros desconhecidos $\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ pode ser solucionado numericamente. O método escore de Fisher é usado para obter as estimativas de máxima verossimilhança de $\hat{\boldsymbol{\beta}}$.

Suponha que os dados são provenientes de um estudo de seleção do habitat, e que as $n_1$ unidades são todas observadas e usadas, mas as $n_2$ unidades são uma amostra aleatória grande de um número muito maior de $N - n_1$ unidades não registradas como sendo usadas. Então, o segundo termo na função de verossimilhança (3.6) pode ser usado para estimar a contribuição média para a verossimilhança para uma unidade não usada, a qual é,

$$Q_{\text{média}} = \left\{ \sum_{i=n_1+1}^{n_1+n_2} \log \frac{(1 - P_i)}{n_2} \right\}.$$

Multiplicando esta média pelo número total de unidades não usadas em toda a área de estudo, têm-se uma estimativa da contribuição do logaritmo da verossimilhança para todas as unidades não usadas. Assim, o logaritmo da pseudo-verossimilhança para todas as unidades dentro da população é:

$$\log(L^*) = \sum_{i=1}^{n_1} \log(P_i) + (N - n_1) Q_{\text{média}} \tag{3.8}$$

Propomos que a RSPF dada pela expressão (3.1) seja estimada pela maximização do logaritmo da pseudo-verossimilhança definida pela expressão (3.8).

Uma aproximação similar foi sugerida por McCracken et al., (1998) com a modelagem de escolha discreta, em que uma unidade é escolhida para o uso de um número grande de unidades, sendo interessante estimar a probabilidade desta unidade particular que é escolhida como a função de valores de *p* variáveis medidas na unidade. Essencialmente, a contribuição da função de verossimilhança por todas as unidades que podem ser escolhidas é aproximada usando grandes amostras destas unidades.

O método proposto para estimar a RSPF logística é efetivamente que, para cada peso da unidade de recurso na amostra de $n_2$ unidades não usadas, temos $\frac{(N-n_1)}{n_2}$ para todas as unidades usadas. Isto significa que a estimativa da máxima pseudo-verossimilhança baseada na expressão (3.8) pode ser obtida de um programa padrão para a regressão logística. A maioria destes programas requer que cada ponto dos dados seja definido como o número de ensaios (*m*) e o número de sucessos (*r*) daqueles ensaios. Usualmente, quando estimamos a RSPF pela regressão logística: *m*=1, para ambas unidades usadas e não usadas, *r*=1 para unidades usadas e *r*=0 para uma unidade não usada. Para o método da pseudo-máxima verossimilhança a estimativa de RSPF pode ser obtida fixando *m*=1 para todas as unidades usadas e $m = \frac{(N-n_1)}{n_2}$ para as unidades não usadas. Isto requer que $\frac{(N-n_1)}{n_2}$ seja um número inteiro, o qual pode ser obtido uma vez que *N* e $n_1$ são conhecidos. Por exemplo, supõe-se $n_1 = 80$ e $N = 100.000$. Então $N - n_1 = 99.920$ poderia ser múltiplo de $n_2$, tais como 2.498, 4.996, 9.992 etc.

Collet (1991), mostrou como os dados para o estudo de um caso controle podem ser analisados usando o modelo logístico linear para explorar o efeito de covariáveis particularmente expostas no caso controle. O autor propõe o delineamento para o estudo do método de caso controle, no qual cada pessoa doente é incluída no estudo como um caso a ser comparado com uma ou mais pessoas sadias que serão incluídas como controles. Desta maneira, os casos podem ser comparados com os controles em áreas residenciais ou lugares do trabalho, levando em conta a mistura das variáveis que não são facilmente quantificáveis. Esta foi a primeira aplicação encontrada na literatura, nós propomos uma aplicação médica para desenvolver uma alternativa para o método caso controle logístico usando uma nova estimativa da máxima pseudo-verossimilhança.

Keating e Cherry, (2004), citaram a importância da regressão logística em estudos de seleção de habitat. Mas, em geral, a regressão logística é inapropriada para modelar a seleção do habitat em estudos de uso, e disponibilidade. Apresentaram que a RSF geralmente é usada para o modelo exponencial sendo proporcional à função discriminante logística. Descreveram uma alternativa de um modelo baseado em Lancaster e Imbens (1996), e este método serve para estimar a probabilidade condicional de uso em estudos de uso, e disponibilidade de recursos. Também expressam a utilidade da RSPF reconhecendo a diferença entre o delineamento do estudo disponível e o delineamento do estudo do caso controle. Lancaster e Imbens (1996), no contexto da seleção do recurso, afirmaram que uma amostra aleatória de unidades de recurso disponíveis contém unidades usadas e não usadas. Ademais, o procedimento de estimação de (MANLY et al., 2002) pode potencialmente conduzir a estimadores viesados.

Johnson et al. (2006), em resposta a Keating e Cherry, (2004) argumentaram que o delineamento do estudo da disponibilidade de uso é formulado propriamente em termos de distribuição ponderada, Patil e Rao (1978).

Johnson et al. (2006), informaram que é apropriado usar a regressão logística para estimar a RSF para

o delineamento baseado na disponibilidade de uso de amostras independentes para o uso de unidades de recurso disponíveis. Ele conclui que o método da nova validação para avaliar o comportamento previsto da RSF avaliado no modelo, deve ser proporcional à probabilidade de uso na unidade de recurso.

A proposta do artigo Lele e Keim (2006), é estender as idéias de Johnson et al. (2006) e demonstram que outras formas paramétricas para a RSPF exponencial, permitem a estimativa da probabilidade absoluta. Este método proporciona flexibilidade no modelamento e gera probabilidades absolutas como opostas à probabilidade relativa. Estimativas da probabilidade absoluta sobre o delineamento amostral da disponibilidade de uso poderia ter uma maior vantagem nas análises dos dados das coleiras de cabras nas montanhas do Canadá monitoradas por rádio em ecologia. Eles descreveram em detalhe o método de estimação da máxima verossimilhança simulada (CASELLA; BERGER, 2002) para a estimação dos parâmetros para qualquer RSPF removendo a restrição de somente empregar a RSPF exponencial nas análises de dados da disponibilidade de uso. Fornecendo estimativas e intervalos de confiança para os parâmetros, mostrando que a RSPF logística obteve um melhor ajuste para os dados que a RSPF exponencial. Argumentaram que o modelo logístico não só ajusta melhor os dados mas também proporciona uma probabilidade absoluta de uso mais adequada que a probabilidade relativa de uso.

### 3.2.3 Bootstrapping para a estimação do viés e da variância

Devido ao tamanho da amostra de unidades não usadas ser artificialmente inflacionado com o método da máxima pseudo-verossimilhança da expressão (3.8), as propriedades dos estimadores obtidos, tais como os erros padrões dos parâmetros da regressão, não são aqueles fornecidos pela teoria padrão da regressão logística. Por isso, propomos o uso de técnicas de bootstrapping para avaliar as propriedades dos estimadores de máxima pseudo-verossimilhança.

Foram consideradas três aproximações para a reamostragem bootstrap dos dados, as quais são descritas em detalhes a seguir.

Método 1 bootstrapping paramétrico: O modelo ajustado fornece uma probabilidade estimada de ser observada uma unidade a ser usada, para todas as unidades amostradas, considerando que as $n_1$ unidades observadas e usadas, são todas as unidades que são observadas e usadas enquanto que as $n_2$ unidades não observadas como sendo usadas devem representar as $N-n_1$ unidades não usadas. Dessa forma, o procedimento bootstrap a seguir é usado para gerar novos conjuntos de dados originais.

(a) Gera-se um número aleatório entre 0 e 1. Se este número for menor do que $\frac{n_1}{N}$, então uma das $n_1$ unidades usadas é aleatoriamente selecionada, caso contrário uma das $n_2$ unidades não usadas é selecionada.

(b) A probabilidade da unidade que está sendo observada ser usada, é calculada para a unidade selecionada usando a regressão logística ajustada para os dados originais. Isto é, a unidade é registrada como usada com esta probabilidade, ou caso contrário, registrada como não usada.

(c) Se a unidade é registrada como usada no passo (b) e ocorrem menos do que $n_1$ unidades usadas, a unidade com estes *X* valores juntos é a amostra de unidades observadas a serem usadas. Se já são $n_1$ unidades usadas, então a unidade é descartada. Similarmente, se as unidades registradas como não usada do passo (b) são menores que $n_2$ unidades não usadas, então a unidade com *X* valores juntos é a amostra de unidades não observadas a serem usadas. Se, já são $n_2$ unidades não observadas a serem usadas, então a unidade é descartada.

(d) Os passos (a) a (c) são repetidos *N* vezes, em que *N* é o número total de unidades de recurso na população.

Os resultados deste procedimento são as $n_2$ unidades não usadas, na amostra de bootstrap. Sendo o mesmo nos dados originais, na suposição que $n_2$ foi um número fixo para estes dados. O número de unidades usadas é entretanto uma variável aleatória para cada conjunto bootstrap de dados. Uma vez que o conjunto de dados é gerado, isto é, analisado da mesma maneira como nos dados originais.

Método bootstrapping não paramétrico: este é mais simples e rápido do que o método 1, e envolve a reamostragem de $n_1$ unidades usadas dentro dos dados originais com reposição, para obter a amostra bootstrap de $n_1$ unidades, e a reamostragem similar das $n_2$ unidades não usadas, com reposição, para obter a amostra bootstrap de $n_2$ unidades não usadas. Uma vez que o conjunto bootstrap de dados é gerado, ele é analisado do mesmo modo que os dados originais.

Método 3 bootstrapping paramétrico modificado: é similar ao bootstrapping para métrico mas os passos do método 1, (c) e (d) são alterados da seguinte forma:
(c1) Se a unidade é registrada como usada no passo (b) e estes são menores do que as $n_1$ unidades usadas, a unidade com X valores na amostra de unidades são observadas a serem usadas. Se já são $n_1$ unidades usadas a unidade é descartada. Similarmente, se a unidade é registrada como não usada no passo (b), e são menos que $n_2$ unidades não usadas, então a unidade com *X* valores na amostra de unidades não observadas a serem usadas. Se são $n_2$ unidades não são observadas a serem usadas então a unidade é descartada.

(d1) Se o número de unidades observadas a serem usadas é menor do que $n_1$, ou a amostra de unidades não observadas como sendo usadas é menor do que $n_2$, então retornamos ao passo (a), caso contrário deve-se parar.

O procedimento acima finaliza com o tamanho da amostra de unidades usadas e não usadas sendo o mesmo que para os dados originais. Os dados bootstrap obtidos são analisados exatamente do mesmo modo como os dados originais.

### 3.2.4 Estudo de simulação e reamostragem de Bootstrap

Realizou-se um estudo de simulação para examinar as propriedades dos coeficientes da regressão estimada pelo método da máxima- pseudo verossimilhança bem como a eficácia do método bootstrapping. Todas as simulações foram baseadas em um modelo de população artificial consistindo em um milhão de unidades de recurso, com cada unidade tendo valores para três variáveis, e com cada variável *X* sendo distribuída normalmente com média zero e variância 1.

O modelo de população da RSPF arbitrário é dado por,

$$w^*(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{\exp(\boldsymbol{\beta}_0 + 2.0x_1 + 1.0x_2 + 0.0x_3)}{1 + \exp(\boldsymbol{\beta}_0 + 2.0x_1 + 1.0x_2 + 0.0x_3)}.$$

Assim, $X_3$ não é importante para a seleção do recurso e o valor de $\boldsymbol{\beta}_0$ fixa a probabilidade geral da unidade ser registrada como sendo usada por um animal. A figura (3.1) mostra a função para o caso particular em que um milhão de unidades de recurso são esperadas para serem usadas.

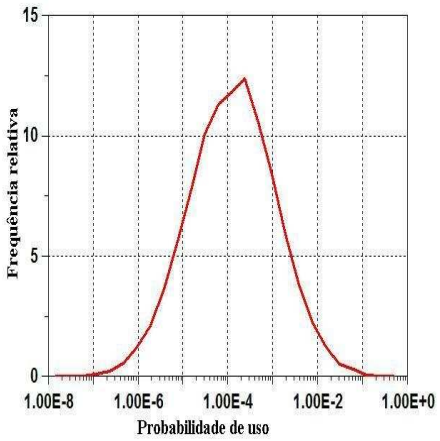



Figura 3.1 – Probabilidade relativa de uso para grandes amostras de unidades de recurso quando o número total esperado de unidades usadas é 1000 ou 0,1% de todas as unidades. Probabilidades de uso variam de $10^{-8}$ a 0,25 com a média de 0,001

Para o estudo de simulação o número esperado de unidades usadas dentro da população $(n_1)$ foi o conjunto de 50, 100, 200, 400 e 800. Para cada um destes valores esperados $n_1$, os números de unidades não usadas $(n_2)$ foi o conjunto de 1.000, 5.000, 25.000. Assim, são $5 \times 3 = 15$ diferentes cenários considerados. Para cada cenário foram gerados 100 conjuntos bootstrap de dados, nos quais foram aplicados os três métodos bootstrap descritos acima. Para dados reais, mais do que 100 conjuntos bootstrap de dados deveriam ser usados. Se por alguma razão os resultados bootstrap obtidos durante o estudo de simulação não forem bons, então as análises podem ser feitas com dados reais.

Os resultados do estudo bootstrap paramétrico foram resumidos desta maneira, a tabela (3.1), apresenta os resultados do bootstrapping paramétrico, e tabela (3.2) com os resultados do método de bootstrapping não paramétrico, usando 1.000, 5.000 e 25.000 unidades não usadas, e apresentando os coeficientes de regressão verdadeiros, as médias e os desvios padrão para os 100 conjuntos de dados gerados, o viés aparente nas estimativas, o viés da média estimada e as médias dos desvios padrão estimado pelo bootstrapping. Também observa-se que as médias e as estimativas do bootstrap dos desvios padrão são próximas aos desvios padrão observados, mostrando que o bootstrap trabalha.

A tabela (3.3) do bootstrapping paramétrico modificado mostra os coeficientes de regressão verdadeiros, as médias e os desvios padrão para os 100 conjuntos de dados gerados, o viés aparente nas estimativas, o viés da média estimada e as médias dos desvios padrão estimado pelo bootstrapping. Estes resultados do bootstrapping paramétrico são apresentados por obter os melhores resultados com o uso do bootstrapping paramétrico do que o bootstrapping não paramétrico.

Também, a tabela (3.3) mostra que os valores estimados da constante de regressão $\boldsymbol{\beta}_0$ têm viés negativo, e este diminui como o aumento do número de unidades não usadas e aumenta quando o tamanho da amostra para as unidades usadas aumenta. Em contraste a isso, as estimativas dos coeficientes de regressão não nulas para $\boldsymbol{\beta}_1$ e $\boldsymbol{\beta}_2$ tem viés positivos, e este também diminui à medida que o número de unidades usadas aumenta e o tamanho da amostra para unidades não usadas aumenta. De qualquer forma, nota-se que o viés é a estimativa do coeficiente nulo $\boldsymbol{\beta}_3$. Esta pequena amostra de viés é provavelmente devido à seleção parecer mais extrema do que realmente é.

Tabela 3.1 – Resultados do primeiro método do bootstrapping paramétrico, usando 1.000, 5.000 e 25.000 unidades não usadas apresentado os coeficientes de regressão verdadeiros as respectivas médias, desvios padrão, viés aparente nas estimativas, o viés da média estimada e as médias dos desvios padrão estimado pelo bootstrapping para 100 conjuntos de dados gerados

| Número aproximado | | 1.000 unidades não usadas | | | | | | 5.000 unidades não usadas | | | | | | 25.000 unidades não usadas | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $Estimativas^c$ | | | | $Bootstrap^d$ | | $Estimativas^c$ | | | | $Bootstrap^d$ | | $Estimativas^c$ | | | | $Bootstrap^d$ | |
| $Usado^a$ | Parâmetros[b] | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão |
| 50 | $\beta_0$ | 12,48 | -13,44 | 1,34 | -0,85 | -0,78 | 1,42 | -12,33 | -12,52 | 0,49 | -0,19 | -0,21 | 0,55 | -12,33 | -12,40 | 0,39 | -0,07 | -0,11 | 0,42 |
| | $\beta_1$ | 2,00 | 2,39 | 0,66 | 0,39 | 0,32 | 0,65 | 2,00 | 2,06 | 0,23 | 0,06 | 0,07 | 0,25 | 2,00 | 2,00 | 0,18 | 0,00 | 0,03 | 0,19 |
| | $\beta_2$ | 1,00 | 1,17 | 0,37 | 0,17 | 0,16 | 0,42 | 1,00 | 1,07 | 0,19 | 0,07 | 0,04 | 0,21 | 1,00 | 1,03 | 0,16 | 0,03 | 0,02 | 0,16 |
| | $\beta_3$ | 0,00 | 0,06 | 0,31 | 0,06 | 0,01 | 0,31 | 0,00 | 0,02 | 0,18 | 0,02 | 0,00 | 0,18 | 0,00 | 0,01 | 0,14 | 0,01 | 0,00 | 0,15 |
| 100 | $\beta_0$ | 11,77 | -12,43 | 0,95 | -0,66 | -0,59 | 0,96 | -11,66 | -11,85 | 0,38 | -0,19 | -0,16 | 0,42 | -11,66 | -11,73 | 0,28 | -0,07 | -0,08 | 0,31 |
| | $\beta_1$ | 2,00 | 2,34 | 0,50 | 0,34 | 0,26 | 0,48 | 2,00 | 2,07 | 0,19 | 0,07 | 0,06 | 0,21 | 2,00 | 2,01 | 0,12 | 0,01 | 0,02 | 0,14 |
| | $\beta_2$ | 1,00 | 1,12 | 0,34 | 0,12 | 0,14 | 0,33 | 1,00 | 1,06 | 0,16 | 0,06 | 0,03 | 0,16 | 1,00 | 1,02 | 0,12 | 0,02 | 0,01 | 0,12 |
| | $\beta_3$ | 0,00 | 0,02 | 0,25 | 0,02 | 0,00 | 0,26 | 0,00 | 0,02 | 0,14 | 0,02 | 0,00 | 0,14 | 0,00 | 0,01 | 0,11 | 0,01 | 0,00 | 0,11 |
| 200 | $\beta_0$ | 11,45 | -12,08 | 0,65 | -0,62 | -0,54 | 0,91 | -10,90 | -11,03 | 0,33 | -0,13 | -0,13 | 0,33 | -10,90 | -10,94 | 0,24 | -0,04 | -0,06 | 0,22 |
| | $\beta_1$ | 2,00 | 2,30 | 0,34 | 0,30 | 0,25 | 0,46 | 2,00 | 2,05 | 0,18 | 0,05 | 0,06 | 0,17 | 2,00 | 2,01 | 0,11 | 0,01 | 0,02 | 0,12 |
| | $\beta_2$ | 1,00 | 1,18 | 0,31 | 0,18 | 0,13 | 0,31 | 1,00 | 1,05 | 0,14 | 0,05 | 0,03 | 0,13 | 1,00 | 1,02 | 0,10 | 0,02 | 0,01 | 0,09 |
| | $\beta_3$ | 0,00 | 0,01 | 0,25 | 0,01 | 0,00 | 0,24 | 0,00 | 0,02 | 0,12 | 0,02 | 0,00 | 0,11 | 0,00 | 0,00 | 0,09 | 0,00 | 0,00 | 0,08 |
| 400 | $\beta_0$ | 10,36 | -10,85 | 0,60 | -0,49 | -0,40 | 0,62 | -10,21 | -10,32 | 0,29 | -0,11 | -0,11 | 0,28 | -10,21 | -10,25 | 0,19 | -0,04 | -0,05 | 0,17 |
| | $\beta_1$ | 2,00 | 2,27 | 0,35 | 0,27 | 0,20 | 0,35 | 2,00 | 2,05 | 0,16 | 0,05 | 0,05 | 0,15 | 2,00 | 2,01 | 0,09 | 0,01 | 0,02 | 0,09 |
| | $\beta_2$ | 1,00 | 1,10 | 0,34 | 0,25 | 0,10 | 0,25 | 1,00 | 1,04 | 0,12 | 0,04 | 0,03 | 0,12 | 1,00 | 1,02 | 0,07 | 0,02 | 0,01 | 0,07 |
| | $\beta_3$ | 0,00 | -0,01 | 0,21 | -0,01 | 0,00 | 0,20 | 0,00 | 0,02 | 0,10 | 0,02 | 0,00 | 0,10 | 0,00 | 0,01 | 0,07 | 0,01 | 0,00 | 0,06 |
| 800 | $\beta_0$ | -9,56 | -9,84 | 0,46 | -0,28 | -0,32 | 0,51 | -9,56 | -9,65 | 0,24 | -0,09 | -0,09 | 0,24 | -9,56 | -9,59 | 0,15 | -0,04 | -0,04 | 0,14 |
| | $\beta_1$ | 2,00 | 2,13 | 0,28 | 0,13 | 0,16 | 0,30 | 2,00 | 2,04 | 0,14 | 0,04 | 0,04 | 0,14 | 2,00 | 2,01 | 0,08 | 0,01 | 0,02 | 0,07 |
| | $\beta_2$ | 1,00 | 1,09 | 0,22 | 0,09 | 0,09 | 0,22 | 1,00 | 1,03 | 0,10 | 0,03 | 0,02 | 0,10 | 1,00 | 1,01 | 0,06 | 0,01 | 0,01 | 0,06 |
| | $\beta_3$ | 0,00 | -0,02 | 0,19 | -0,02 | 0,00 | 0,17 | 0,00 | 0,02 | 0,09 | 0,02 | 0,00 | 0,09 | 0,00 | 0,01 | 0,05 | 0,01 | 0,00 | 0,05 |

[a] Os Dados gerados usando as probabilidades de sendo observadas a ser usadas de tal modo que esse número observado atual a ser usado é a variável aleatória com o valor esperado apresentado.

[b] O valor do parâmetro verdadeiro usado para gerar dado, o qual são os mesmos para os três tamanhos de amostras de unidades usadas.

[c] As médias e os desvios padrão e o viés da média (estimativa da média- valor verdadeiro) de estimativas de 100 conjuntos de dados gerados.

[d] A média da estimativa do viés e a média da estimativa do desvio padrão do bootstrapping paramétrico aplicado a cada 100 conjuntos de dados gerados com 100 conjunto bootstrap de dados para cada conjunto gerado de dados.

Tabela 3.2 – Resultados do segundo método do bootstrapping não paramétrico, usando 1.000, 5.000 e 25.000 unidades não usadas apresentado os coeficientes de regressão verdadeiros as respectivas médias, desvios padrão, viés aparente nas estimativas, o viés da média estimada e as médias dos desvios padrão estimado pelo bootstrapping para 100 conjuntos de dados gerados

| Número aproximado | | 1.000 unidades não usadas | | | | | | 5.000 unidades não usadas | | | | | | 25.000 unidades não usadas | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $Estimativas^c$ | | | | $Bootstrap^d$ | | $Estimativas^c$ | | | | $Bootstrap^d$ | | $Estimativas^c$ | | | | $Bootstrap^d$ | |
| $Usado^a$ | Parâmetros[b] | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Valor verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão |
| 50 | $\beta_0$ | 12,48 | -13,28 | 1,18 | -0,80 | -0,52 | 1,19 | -12,48 | -12,72 | 0,56 | -0,24 | -0,19 | 0,60 | -12,33 | -12,48 | 0,40 | -0,15 | -0,08 | 0,41 |
| | $\beta_1$ | 2,00 | 2,36 | 0,59 | 0,36 | 0,22 | 0,56 | 2,00 | 2,08 | 0,27 | 0,08 | 0,07 | 0,28 | 2,00 | 2,06 | 0,18 | 0,06 | 0,02 | 0,19 |
| | $\beta_2$ | 1,00 | 1,19 | 0,37 | 0,19 | 0,12 | 0,39 | 1,00 | 1,05 | 0,22 | 0,05 | 0,04 | 0,22 | 1,00 | 1,02 | 0,15 | 0,02 | 0,01 | 0,16 |
| | $\beta_3$ | 0,00 | 0,00 | 0,33 | 0,00 | -0,01 | 0,31 | 0,00 | -0,01 | 0,21 | -0,01 | 0,00 | 0,19 | 0,00 | -0,01 | 0,15 | -0,01 | 0,00 | 0,15 |
| 100 | $\beta_0$ | -11,77 | -12,21 | 0,66 | -0,44 | -0,35 | 0,77 | -11,77 | -11,99 | 0,43 | -0,22 | -0,12 | 0,42 | -11,66 | -11,71 | 0,31 | -0,05 | -0,04 | 0,29 |
| | $\beta_1$ | 2,00 | 2,20 | 0,37 | 0,20 | 0,17 | 0,40 | 2,00 | 2,08 | 0,22 | 0,08 | 0,05 | 0,21 | 2,00 | 2,02 | 0,15 | 0,02 | 0,01 | 0,14 |
| | $\beta_2$ | 1,00 | 1,11 | 0,26 | 0,11 | 0,09 | 0,29 | 1,00 | 1,06 | 0,15 | 0,06 | 0,03 | 0,17 | 1,00 | 1,01 | 0,12 | 0,01 | 0,01 | 0,12 |
| | $\beta_3$ | 0,00 | 0,02 | 0,25 | 0,02 | 0,01 | 0,24 | 0,00 | 0,01 | 0,14 | 0,01 | 0,00 | 0,15 | 0,00 | 0,02 | 0,13 | 0,02 | 0,00 | 0,11 |
| 200 | $\beta_0$ | -10,98 | -11,52 | 0,75 | -0,54 | -0,28 | 0,61 | -10,98 | -11,09 | 0,35 | -0,11 | -0,09 | 0,31 | -10,90 | -10,96 | 0,23 | -0,06 | -0,03 | 0,20 |
| | $\beta_1$ | 2,00 | 2,24 | 0,39 | 0,24 | 0,14 | 0,34 | 2,00 | 2,04 | 0,18 | 0,04 | 0,04 | 0,17 | 2,00 | 2,02 | 0,12 | 0,02 | 0,01 | 0,10 |
| | $\beta_2$ | 1,00 | 1,18 | 0,31 | 0,18 | 0,08 | 0,25 | 1,00 | 1,01 | 0,15 | 0,01 | 0,02 | 0,13 | 1,00 | 1,03 | 0,10 | 0,03 | 0,00 | 0,09 |
| | $\beta_3$ | 0,00 | 0,03 | 0,22 | 0,03 | 0,00 | 0,21 | 0,00 | 0,02 | 0,11 | 0,02 | 0,00 | 0,12 | 0,00 | 0,01 | 0,08 | 0,01 | 0,00 | 0,08 |
| 400 | $\beta_0$ | -10,36 | -10,80 | 0,57 | -0,44 | -0,23 | 0,51 | -10,36 | -10,48 | 0,27 | -0,12 | -0,08 | 0,27 | -10,21 | -10,27 | 0,17 | -0,05 | -0,03 | 0,16 |
| | $\beta_1$ | 2,00 | 2,21 | 0,34 | 0,21 | 0,12 | 0,30 | 2,00 | 2,05 | 0,16 | 0,05 | 0,04 | 0,15 | 2,00 | 2,02 | 0,09 | 0,02 | 0,01 | 0,08 |
| | $\beta_2$ | 1,00 | 1,14 | 0,25 | 0,14 | 0,06 | 0,22 | 1,00 | 1,03 | 0,13 | 0,03 | 0,02 | 0,12 | 1,00 | 1,02 | 0,07 | 0,02 | 0,01 | 0,07 |
| | $\beta_3$ | 0,00 | 0,02 | 0,21 | 0,02 | 0,00 | 0,19 | 0,00 | -0,01 | 0,10 | -0,01 | 0,00 | 0,10 | 0,00 | -0,01 | 0,07 | -0,01 | 0,00 | 0,06 |
| 800 | $\beta_0$ | -9,66 | -10,07 | 0,48 | -0,42 | -0,19 | 0,43 | -9,66 | -9,78 | 0,27 | -0,12 | -0,06 | 0,22 | -9,56 | -9,59 | 0,14 | -0,03 | -0,02 | 0,13 |
| | $\beta_1$ | 2,00 | 2,20 | 0,30 | 0,20 | 0,11 | 0,27 | 2,00 | 2,06 | 0,15 | 0,06 | 0,03 | 0,13 | 2,00 | 2,02 | 0,07 | 0,02 | 0,01 | 0,07 |
| | $\beta_2$ | 1,00 | 1,14 | 0,19 | 0,14 | 0,05 | 0,19 | 1,00 | 1,03 | 0,11 | 0,03 | 0,01 | 0,10 | 1,00 | 1,01 | 0,06 | 0,01 | 0,00 | 0,06 |
| | $\beta_3$ | 0,00 | 0,00 | 0,19 | 0,00 | 0,00 | 0,17 | 0,00 | 0,01 | 0,09 | 0,01 | 0,00 | 0,08 | 0,00 | 0,01 | 0,06 | 0,01 | 0,00 | 0,05 |

[a] Os Dados gerados usando as probabilidades de sendo observadas a ser usadas de tal modo que esse número observado atual a ser usado é a variável aleatória com o valor esperado apresentado.

[b] O valor do parâmetro verdadeiro usado para gerar dado, o qual são os mesmos para os três tamanhos de amostras de unidades usadas.

[c] As médias e os desvios padrão e o viés da média (estimativa da média- valor verdadeiro) de estimativas de 100 conjuntos de dados gerados.

[d] A média da estimativa do viés e a média da estimativa do desvio padrão do bootstrapping paramétrico aplicado a cada 100 conjuntos de dados gerados com 100 conjunto bootstrap de dados para cada conjunto gerado de dados.

Tabela 3.3 – Resultados do terceiro método do bootstrapping paramétrico modificado, usando 1.000, 5.000 e 25.000 unidades não usadas apresentado os coeficientes de regressão verdadeiros as respectivas médias, desvios padrão, viés aparente nas estimativas, o viés da média estimada e as médias dos desvios padrão estimado pelo bootstrapping para 100 conjuntos de dados gerados

| | | | 1.000 unidades não usadas | | | | | 5.000 unidades não usadas | | | | | 25.000 unidades não usadas | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Número aproximado | | | $Estimativas^c$ | | | $Bootstrap^d$ | | Estimativas | | | Bootstrap | | Estimativas | | | Bootstrap | |
| $Usado^a$ | Parâmetros$^b$ | Valor Verdadeiro | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão | Média | Desvios Padrão | Viés observado | Viés | Desvios Padrão |
| 50 | $\beta_0$ | -12,33 | -12,94 | 0,94 | -0,61 | -0,70 | 1,28 | -12,55 | 0,54 | -0,22 | -0,19 | 0,52 | -12,44 | 0,34 | -0,11 | -0,09 | 0,39 |
| | $\beta_1$ | 2,0 | 2,26 | 0,45 | 0,26 | 0,30 | 0,60 | 2,07 | 0,26 | 0,07 | 0,07 | 0,25 | 2,03 | 0,14 | 0,03 | 0,03 | 0,18 |
| | $\beta_2$ | 1,0 | 1,15 | 0,38 | 0,15 | 0,16 | 0,40 | 1,07 | 0,21 | 0,07 | 0,04 | 0,20 | 1,03 | 0,15 | 0,03 | 0,02 | 0,16 |
| | $\beta_3$ | 0,0 | 0,04 | 0,21 | 0,04 | 0,00 | 0,29 | -0,01 | 0,19 | -0,01 | 0,00 | 0,20 | 0,00 | 0,14 | 0,00 | 0,00 | 0,15 |
| 100 | $\beta_0$ | -11,66 | -12,17 | 0,74 | -0,51 | -0,51 | 0,88 | -11,80 | 0,40 | -0,14 | -0,15 | 0,39 | -11,74 | 0,30 | -0,08 | -0,07 | 0,29 |
| | $\beta_1$ | 2,0 | 2,25 | 0,41 | 0,25 | 0,24 | 0,46 | 2,04 | 0,21 | 0,04 | 0,06 | 0,20 | 2,02 | 0,14 | 0,02 | 0,03 | 0,14 |
| | $\beta_2$ | 1,0 | 1,10 | 0,30 | 0,10 | 0,13 | 0,31 | 1,05 | 0,16 | 0,05 | 0,03 | 0,16 | 1,04 | 0,13 | 0,04 | 0,02 | 0,12 |
| | $\beta_3$ | 0,0 | 0,0 | 0,28 | 0,00 | 0,00 | 0,24 | 0,02 | 0,16 | 0,02 | 0,00 | 0,14 | 0,00 | 0,10 | 0,00 | 0,00 | 0,11 |
| 200 | $\beta_0$ | -10,90 | -11,39 | 0,62 | -0,49 | -0,44 | 0,74 | -11,12 | 0,37 | -0,21 | -0,13 | 0,33 | -10,95 | 0,26 | -0,04 | -0,06 | 0,21 |
| | $\beta_1$ | 2,00 | 2,25 | 0,34 | 0,25 | 0,21 | 0,40 | 2,10 | 0,19 | 0,10 | 0,06 | 0,17 | 2,02 | 0,13 | 0,02 | 0,02 | 0,11 |
| | $\beta_2$ | 1,00 | 1,12 | 0,26 | 0,12 | 0,11 | 0,28 | 1,05 | 0,13 | 0,05 | 0,03 | 0,13 | 1,02 | 0,09 | 0,02 | 0,01 | 0,09 |
| | $\beta_3$ | 0,00 | 0,00 | 0,19 | 0,00 | 0,00 | 0,21 | 0,02 | 0,13 | 0,02 | 0,00 | 0,12 | -0,01 | 0,08 | 0,01 | 0,00 | 0,08 |
| 400 | $\beta_0$ | -10,21 | -10,67 | 0,61 | -0,46 | -0,38 | 0,62 | -10,30 | 0,30 | -0,08 | -0,11 | 0,26 | -10,27 | 0,17 | -0,06 | -0,04 | 0,17 |
| | $\beta_1$ | 2,00 | 2,22 | 0,36 | 0,22 | 0,19 | 0,35 | 2,03 | 0,17 | 0,03 | 0,05 | 0,15 | 2,02 | 0,08 | 0,02 | 0,02 | 0,09 |
| | $\beta_2$ | 1,00 | 1,15 | 0,25 | 0,15 | 0,10 | 0,24 | 1,01 | 0,12 | 0,01 | 0,03 | 0,12 | 1,01 | 0,07 | 0,01 | 0,01 | 0,07 |
| | $\beta_3$ | 0,00 | 0,00 | 0,20 | 0,00 | 0,00 | 0,19 | 0,01 | 0,10 | 0,01 | 0,00 | 0,10 | -0,01 | 0,07 | -0,01 | 0,00 | 0,06 |
| 800 | $\beta_0$ | -9,56 | -9,88 | 0,47 | -0,32 | -0,31 | 0,50 | -9,66 | 0,23 | -0,10 | -0,09 | 0,24 | -9,60 | 0,13 | -0,05 | -0,04 | 0,13 |
| | $\beta_1$ | 2,00 | 2,17 | 0,29 | 0,17 | 0,16 | 0,30 | 2,05 | 0,14 | 0,05 | 0,05 | 0,14 | 2,01 | 0,07 | 0,01 | 0,01 | 0,07 |
| | $\beta_2$ | 1,00 | 1,08 | 0,21 | 0,08 | 0,09 | 0,22 | 1,03 | 0,09 | 0,03 | 0,02 | 0,11 | 1,01 | 0,06 | 0,01 | 0,01 | 0,06 |
| | $\beta_3$ | 0,00 | 0,02 | 0,18 | 0,02 | 0,00 | 0,18 | 0,00 | 0,08 | 0,00 | 0,00 | 0,09 | 0,00 | 0,06 | 0,00 | 0,00 | 0,05 |

$^a$ Os Dados gerados usando as probabilidades de sendo observadas a ser usadas de tal modo que esse número observado atual a ser usado é a variável aleatória com o valor esperado apresentado.

$^b$ O valor do parâmetro verdadeiro usado para gerar dado, o qual são os mesmos para os três tamanhos de amostras de unidades usadas.

$^c$ As médias e os desvios padrão e o viés da média (estimativa da média- valor verdadeiro) de estimativas de 100 conjuntos de dados gerados.

$^d$ A média da estimativa do viés e a média da estimativa do desvio padrão de bootstrapping paramétrico aplicado a cada 100 conjuntos de dados gerados com 100 conjunto bootstrap de dados para cada conjunto gerado de dados.

Isso é uma indicação do viés devido a amostras pequenas durante as estimativas bootstrap do desvio padrão das estimativas dos coeficientes da regressão. O que significa que as estimativas bootstrap dos desvios padrão são usualmente mais altas do que os desvios padrão observados das estimativas dos coeficientes da regressão com a amostra de 1.000 unidades não usadas e 200, ou menos, unidades usadas.

Ainda que os coeficientes de regressão estimados mostrem o viés, parece que eles são melhor estimados pelo bootstrapping paramétrico, ou ao menos o termo médio. Similarmente, o desvio padrão observado dos parâmetros da regressão estimada é próximo à média da estimativa bootstrap quando há mais do que 1.000 unidades não usadas ou mais de que 200 unidades usadas. Os resultados da tabela (3.3) indicam que este bootstrapping é efetivo para estimar e remover o viés dos coeficientes da regressão estimados, para determinar o desvio padrão destes coeficientes, com a condição desta amostra de unidades usadas e da amostra de unidades não usadas, não serem pequenas. Se as estimativas bootstrap de viés e desvios padrão são próximos aos valores exatos, então as inferências com respeito aos coeficientes da regressão podem ser baseadas na suposição de que,

$$Z_i = \frac{b_i - \text{Viés}_i - \widehat{\boldsymbol{\beta}}_i}{DP_i} \tag{3.9}$$

segue aproximadamente a distribuição normal padrão, em que $b_i$ é a estimativa da pseudo-verossimilhança do parâmetro $\boldsymbol{\beta}_i$ na expressão (3.1). $\text{Viés}_i$, é a estimativa do viés bootstrap para a estimativa $\widehat{\boldsymbol{\beta}}_i$, e $DP_i$ é a estimativa bootstrap do desvio padrão de $b_i$. Por exemplo, se a aproximação normal é válida, é possível a construção de um intervalo de confiança de $95\%$ para $\widehat{\boldsymbol{\beta}}_i$ como,

$$b_i - \text{Viés}_i - 1,96(DP_i) < \boldsymbol{\beta}_i < b_i - \text{Viés}_i + 1,96(DP_i)$$

Se $b_i$ for significativamente diferente de zero ao nível de $5\%$, se $\frac{b_i - \text{Viés}_i}{DP_i}$, estiver fora da amplitude de $-1,96$ a $1,96$.

Quando a tabela 3.4 indica esta suposição para os valores de $Z_i$, seguindo a expressão (3.9) para $Z_i$ com média zero e desvio padrão um, é razoável assumir um tamanho da amostra de 5.000 ou 25.000 unidades não usadas, mas não em que o tamanho da amostra é de 1.000. Com somente a amostra de 1.000 unidades não usadas os desvios padrão simulados de Z valores são usualmente também baixos, porque o viés é positivo dentro da estimativa bootstrap do desvio padrão do coeficiente de regressão.

Com 5.000 ou 25.000 unidades não usadas o desvio padrão dos $Z_i$ valores é significativamente diferentes de cinco vezes o nível de $5\%$ mas com um modelo não muito claro, porque três dos cinco desvios padrão, são maiores do que ele e dois são menores do que ele. Há uma pequena evidência de que a média dos valores Z difere de zero, tal que sobre todos valores Z com média de zero e desvio padrão um, parece razoável, se o tamanho da amostra de unidades não usadas é 5.000.

Para inferências é também necessário que estes valores $Z_i$ da expressão (3.9) tenham distribuição, e que sejam próximas da distribuição normal padrão. Em particular $\varphi(Z)$ poderia ter distribuição próxima à uniforme entre zero e um, em que $\varphi(Z)$ é uma função de distribuição normal cumulativa. Este não é o caso com somente 1.000 amostras de unidades não usadas, mas esta é uma boa aproximação para 5.000 e 25.000 unidades não usadas, como mostra a figura (3.2). Neste esquema podem ser notadas que para 1.000 unidades não usadas da porcentagem de $Z_i$ valores significativamente diferente de zero, para o nível de 5% de

significância, os valores sublinhados dos desvios padrão apresentam-se fora da amplitude de 0,86 a 1,4 sendo muito baixo. Já para 5.000 e 25.000 as estimativas das médias e dos desvios padrão de $z_i$ estão próximas do desejado 5%, sendo 5,2% e 4,7% respectivamente.

Na figura (3.2) estão os 2.000 valores Z obtidos pelo estudo de simulação, para as estimativas usando números diferentes para a estimação de unidades não usadas e utilizadas. A distribuição de $\varphi(Z)$ poderia ser uniforme entre zero e um. Contagem de intervalos de largura de 0,025 são apresentados de cor vermelho, junto com as contagens esperadas, se a suposição da distribuição uniforme é correta, cor azul. A suposição é razoável mas não para 1000 unidades não usadas.

Tabela 3.4 – Médias e desvios padrão dos valores $Z_i$ para o estudo de simulação

| Valores médios esperados | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cenários | Unidades não usadas | Unidades usadas | $Z_0$ | | $Z_1$ | | $Z_2$ | | $Z_3$ | |
| | | | Médias | Desvios Padrão | Médias | Desvios Padrão | Médias | Desvios Padrão | Médias | Desvios Padrão |
| 1 | 1000 | 50 | 0,24 | 0,71 | -0,21 | 0,75 | -0,14 | 0,82 | 0,12 | 0,68 |
| 2 | 1000 | 100 | 0,12 | 0,77 | -0,05 | 0,83 | -0,14 | 0,89 | 0,00 | 1,01 |
| 3 | 1000 | 200 | 0,04 | 0,78 | 0,02 | 0,82 | -0,01 | 0,86 | -0,05 | 0,88 |
| 4 | 1000 | 400 | -0,03 | 0,79 | -0,02 | 0,83 | 0,15 | 0,87 | 0,02 | 0,98 |
| 5 | 1000 | 800 | 0,07 | 0,81 | -0,05 | 0,84 | -0,07 | 0,89 | 0,08 | 0,99 |
| 6 | 5000 | 50 | 0,15 | 1,07 | -0,12 | 1,05 | 0,05 | 0,97 | -0,03 | 1,04 |
| 7 | 5000 | 100 | 0,12 | 0,95 | -0,17 | 0,99 | 0,08 | 1,00 | 0,13 | 1,10 |
| 8 | 5000 | 200 | -0,14 | 1,02 | 0,13 | 1,02 | 0,12 | 0,96 | 0,18 | 1,15 |
| 9 | 5000 | 400 | 0,19 | 1,09 | -0,17 | 1,09 | -0,17 | 0,95 | 0,13 | 1,01 |
| 10 | 5000 | 800 | 0,03 | 0,94 | -0,03 | 1,02 | 0,02 | 0,86 | 0,03 | 0,94 |
| 11 | 25000 | 50 | 0,05 | 0,81 | -0,03 | 0,74 | 0,09 | 0,87 | -0,02 | 0,90 |
| 12 | 25000 | 100 | 0,12 | 1,07 | -0,10 | 0,97 | 0,13 | 1,04 | -0,03 | 0,95 |
| 13 | 25000 | 200 | 0,19 | 1,25 | -0,16 | 1,16 | 0,02 | 1,03 | -0,16 | 0,96 |
| 14 | 25000 | 400 | -0,03 | 1,03 | 0,00 | 0,95 | 0,00 | 0,98 | -0,08 | 1,05 |
| 15 | 25000 | 800 | -0,03 | 0,96 | -0,02 | 0,90 | -0,02 | 0,93 | 0,09 | 1,06 |

Médias e desvios padrão dos valores $Z_i$ como definidos pela expressão (3.9) para o estudo de simulação descrito na seção anterior. Os valores sublinhados da média são significativamente diferentes de zero ao nível de 5% (fora da amplitude -0,20 a +0,20). Os valores sublinhados dos desvios padrão são significativamente diferentes de um ao nível de 5% (fora da amplitude 0,86 a 1,14)

Os resultados da simulação anterior do experimento mostram que as inferências sobre os coeficientes de regressão estimados, podem ser baseados na suposição de que estes são distribuídos normalmente como um viés de pequenas amostras, e as variâncias podem ser estimadas razoavelmente bem pela reamostragem bootstrap paramétrica, mostrando que este tamanho de amostras para unidades não usadas é suficientemente grande.

Para as situações consideradas durante as simulações, o tamanho da amostra de 5.000 ou mais unidades não usadas parece ser razoável. Isto é consistente com a conclusão de Nielson et al., (2004) sendo uma justificativa para usar mais do que 1.000 unidades de recurso disponíveis quando a estimativa da RSF usa regressão logística ordinária com a amostra de unidades observadas a serem usadas, e a amostra aleatória da população de unidades de recurso disponíveis.

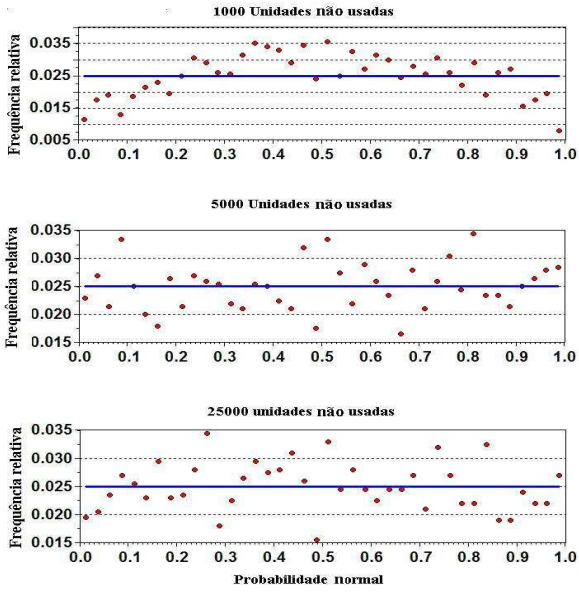



Figura 3.2 – Distribuição dos valores de $\varphi(Z)$ para os valores de Z na expressão (3.9) obtidos no experimento de simulação descrito na seção anterior, em que $\varphi(.)$ é a função de distribuição acumulativa para a distribuição normal padrão

## 3.3 Discussão

Existem inúmeras questões que precisam ser consideradas quando se usa o método da regressão logística descrito neste capítulo para estimar a RSPF . Por exemplo, este método fornece uma estimativa da probabilidade de que a unidade de recurso que é observada é usada. Se, nem todas as unidades usadas são observadas e usadas, então a probabilidade desta ser estimada pode ser interpretada como a probabilidade de unidades a serem usadas, multiplicada pela probabilidade $\boldsymbol{\theta}$ sendo a unidade registrada como usada. Se $\boldsymbol{\theta}$ é o mesmo para todas as unidades, então este pode ser considerado aceitável. De qualquer forma, se $\boldsymbol{\theta}$ varia de acordo com a natureza da unidade de recurso, então a RSPF é viesada como termo da estimativa do uso atual do recurso. Não iremos considerar este assunto aqui, mais é um indicativo que pode ser permitido usar os métodos discutidos por Mackenzie (2006).

Uma limitação na suposição da RSPF logística é que se ela é correta para um tempo de duração, então, ela não seria correta para qualquer outro tempo de duração. Para este exemplo, supõe-se que a probabilidade da unidade de recurso que está sendo registrada como usada ao menos uma vez durante um ano é $w_1^*(x_1, \ldots, x_p)$, quando esta é uma função logística dos valores das covariáveis $x_1$ a $x_p$, que são medidas na unidade. Supõe-se também que a probabilidade da unidade que está sendo registrada como usada, como sendo mínima, uma vez que durante o segundo ano é exatamente a mesma, com a independência entre anos. Então, a probabilidade da unidade que está sendo usada uma vez como mínima durante um período de dois anos é,

$$\begin{aligned} w_2^*(x_1, \ldots, x_p) &= 1 - [1 - w_1^*(x_1, \ldots, x_p)]^2 \\ &= 2w_1^*(x_1, \ldots, x_p) - [w_1^*(x_1, \ldots, x_p)]^2 \end{aligned}$$

ou seja 1 menos a probabilidade de que é não registrada como usada em ambos anos. Esta não é uma função

da regressão logística, mostrando que esta suposição da regressão logística não pode ser incluída para ambos os anos no período da seleção de um ano e o período da seleção de dois anos, a menos que a função não seja a mesma durante ambos os anos.

Ainda que isto seja verdadeiro, a função logística poderia ser uma boa aproximação, se a probabilidade de uso durante um ano é pequena para todas as unidades de recurso. Neste caso o numerador para a função logística é,

$$w_1^*(X_1 = x_1, \ldots, X_p = x_p) = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}{1 + \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)}$$

o numerador poderia ser próximo de zero e, nesse caso, o denominador poderia ser próximo de um. Também $w_1^*(x_1, \ldots, x_p)^2$ poderia ser insignificante com relação a expressão para $w_2^*(x_1, \ldots, x_p)$.

Assim,

$$w_1^*(x_1, \ldots, x_p) \approx \exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p)$$

e

$$w_2^*(x_1, \ldots, x_p) \approx 2\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p) = \exp[\log_e(2) + \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \ldots + \boldsymbol{\beta}_p x_p]$$

em que ambas as funções poderiam ser bem aproximadas pela regressão logística.

O significado é que a suposição da regressão logística para a estimativa da RSPF poderia ser razoável com a condição de que todas as unidades tenham probabilidades pequenas de uso para os períodos de tempo de interesse. Se esta não é a situação, então a regressão logística poderia ainda dar uma aproximação empírica razoável para a verdadeira RSPF . No entanto, isto é importante para avaliar esta aproximação, sendo específica para o uso do período de tempo no estudo. O método de estimação da máxima-pseudo verossimilhança descrito neste capítulo pode ser aplicado com outra forma da função que permita, para diferentes períodos de tempo, a melhor seleção, mas isto ainda não foi pesquisado. No sentido de verificar convergência de Lele e Keim (2006), gerou-se um conjunto de dados para uma população de 100.000 unidades de recurso com 9.812 unidades usadas, ou seja 9,8% usado. Com dados de todas as unidades usadas e uma amostra de 10.000 unidades disponíveis, ainda não apresentou convergência pelo programa de R de Lele e Keim (2006), o qual é uma suspeita de uma suficiente alta proporção de unidades a serem usadas.

Conclui-se que o uso da regressão logística para estimar a RSPFs para um dado ano, quando se descrevem dois anos, a função RSPF da probabilidade de uso não seria adequada. A RSPF pode ser estimada usando qualquer programa de regressão logística, mas as variâncias das estimativas dos parâmetros são pequenas e a qualidade da estatística não é boa. Para superar os problemas do viés do tamanho amostral pequeno, na estimativa do parâmetro, propõe-se reamostragem bootstrap. Utilizou-se três métodos: Bootstrap paramétrico, não paramétrico e paramétrico com modificações.

Os resultados da simulação mostram que, as inferências sobre os coeficientes de regressão estimados podem ser baseados na suposição de que estas são distribuídas normalmente, como um viés de pequenas amostras e as variâncias podem ser estimadas razoavelmente bem pela reamostragem bootstrap paramétrica, mostrando que o tamanho de amostras para unidades não usadas (1.000, 5.000, 25.000) é suficientemente grande.

Este método precisaria ser conhecido pelos usuários das análises de dados para estimar as RSPFs descrita aqui, possuindo aplicações potenciais para muitas outras situações. Por exemplo, se há interesse em entender como a probabilidade de uma doença está relacionada com as condições ambientais dentro de uma cidade, pode-se ter informações sobre as condições de vida para $n_1$ pessoas dentro da cidade registrada como tendo a doença e para uma amostra aleatória grande de $n_2$ pessoas dentro da cidade sem a doença. Regressão logística pode então ser usada para relacionar a probabilidade de ser registrada a doença, com as variáveis registradas para cada pessoa (VERGARA et al., 2008).

**Referências**

CASELLA, G; BERGER, R.L. **Statistical inference**. 2nd ed. Duxbury. 2002. 660 p.

JHONSON, C.J.; NIELSEN, S. E.; MERRILL, E. H.; McDONALD, T.L.; BOYCE, M.S. Resource selection functions based on use-availability data: theoretical motivation and evaluation methods. **Journal of Wildlife Management**, Cheyenne, v. 70, n. 2, p. 347-357, 2006.

COLLET, D. **Modelling binary data**. London: Chapman & Hall, 1991. 369 p.

KEATING, A.K.; CHERRY, S. Use and interpretation of logistic regression in habitat-selection studies. **Journal of Wildlife Management**, Cheyenne, v. 68, n. 4, p. 774-789, 2004.

LANCASTER, T; IMBENS, G. Case-control studies with contaminated controls. **Journal of Econometrics**, Amsterdam, v. 71, p. 145-160, 1996.

LELE, S.R.; KEIM, J.L. Weighted distributions and estimation of resource selection probability functions. **Ecology**, Brooklyn, v. 87, n. 12, p. 3021-3028, 2006.

McDONALD, L.L.; MANLY, B.F.J.; RALEY, C. M. Analyzing foraging and habitat use through selection functions. **Studies in Avion Biology**, Los Angeles, v. 13, p. 325-331, 1990.

MACKENZIE, D.I. Modelling the probability of resource use: the effect of and dealing with detecting a species imperfectly. **Journal of Wildlife Management**, Cheyenne, v. 70, p. 367-74, 2006.

McCRACKEN, M.L.; MANLY, B.F.J.; VANDER HEYDEN, M. The use of discrete-choice models for evaluating resource selection. **Journal of Agricultural, Biological and Environmental Statistics**, Alexandria, v. 3, n. 3, p. 268-279, 1998.

MANLY, B.F.J. **The statistics of natural selection**. New York: Chapman and Hall, 1985. 484 p.

MANLY, B.F.J.; McDONALD, L.L.; THOMAS, D.L.; McDONALD, T.L.; ERICKSON, W.P. **Resource selection by animals**. 2nd ed. London: Kluwer Academic Publishers, 2002. 221 p.

NIELSON, R.; MANLY, B.F.J.; McDONALD, L.L. A preliminary study of the bias and variance when estimating a resource selection function with separate samples of used and available resource units. **In Resource Selection Methods and Applications**, Madison, v. 63, p. 28-34, 2004.

PATIL, G.P ; RAO, C.R. Weighted distributions and size-biased sampling with applications to wildlife populations and human families. **Biometrics**, Washington, v. 34, p. 179-189, 1978.

STRICKLAND, M.D.; McDONALD, L.L. Introduction to the special issue. **Journal of Wildlife Management**, Cheyenne, v. 70, p. 321-323, 2006.

TRAIN, K. **Discrete choice methods with simulation.** Cambridge:Cambridge University Press, 2003. 334 p.

VERGARA, C.S.; DIAS, C.T.S.; MANLY, B.F.J. Bootstrapping for estimation of biases and variances at the resource selection probability function. In: 18º SIMPOSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA (SINAPE)., 2008, Estância de São Pedro. **Resumo**. . . São Pedro: ABE, 2008. p. 51.

# 4 MODELOS DE ESCOLHA DISCRETA NA SELEÇÃO DE RECURSO ALEATÓRIO NA SELEÇÃO DE HABITAT DA CORUJA MANCHADA (*Strix occidentalis*)


## Resumo

Neste capítulo estuda-se o comportamento de um grupo de Corujas Manchadas (*Strix occidentalis*) habitando uma região do noroeste da California (USA) e predize-se seu comportamento em relação à seleção de hábitat. As estimativas do comportamento dos animais individuais difere das estimativas dos grupos de animais. McDonald et al. (1990), apresentam a função da probabilidade da seleção do recurso (RSPF ), definida como a função que fornece as probabilidades de uso para as unidades de recurso de diferentes tipos. As idéias apresentadas por McFadden (TRAIN, 2003), originalmente usadas na área econômica, são aplicadas aqui na estimação da RSPF . Os animais fazem as escolhas do alimento presente no habitat a partir de um conjunto de múltiplas escolhas.
Apresentamos aqui a estrutura do modelo encaixado para $B_1, \ldots, B_k$ encaixes não sobrepostos, com $l = 1, \ldots, k, \ldots, L$, alternativas dentro de cada encaixe. As estimativas foram obtidas usando o modelo logit encaixado, o valor extremo heterocedástico e a maximização da utilidade aleatória, através de métodos de otimização numérica.


Palavras-chave: Função da seleção do recurso (RSPF ); maximização da utilidade aleatória; Métodos de otimização numérica; múltiplas escolhas; Modelo logit encaixado


## Abstract

This chapter presents behavioral observations and predictions of a group of spotted owls (*Strix occidentalis*), living in a northwestern region of California state (USA). Behavioral estimates of individual animals differ from estimates of the animal group. McDonald et al. (1990) defined resource selection probability function (RSPF) as the function that provides the probability of use for different resources. The ideas presented by McFadden (TRAIN, 2003), originally applied to economical situations, are used here to estimate the RSPF. Animals make choices of food found in the habitat from a group of multiple choices. We present the structure of a nested model for $B_1, B_2, \ldots, B_k$ non-overlapping nests, with $l = 1, \ldots, L$ alternatives within each nest. Estimates were obtained using the nested logit model, heterocedastic extreme value, and random utility maximization through numerical optimization methods.


Keywords: Resource selection probability function (RSPF); Random utility maximization; Methods of numerical optimization; Multiple choices; Nested logit model

## 4.1 Introdução

Em ecologia, estuda-se o comportamento dos grupos de animais, para assim predizer as estimativas nos comportamentos deles através de modelos agregados. Procura-se estudar o comportamento individual destes para conhecer suas preferências, usando modelos desagregados. Para saber as preferências dos animais com relação à vegetação, que são covariáveis, pode-se usar a idéia da teoria microeconômica e assim, observar as decisões dos grupos de animais, em um conjunto de hipóteses correspondentes a estas preferências. Podemos analisar e interpretar um modelo desagregado, não sendo possível usar este modelo quando estudamos o comportamento dos grupos de animais que utilizam os modelos agregados.

Sendo assim, os modelos desagregados são aptos para efeitos da captura, não podendo ser incorporados com precisão dentro dos modelos agregados, precisando assim, de novas metodologias. Para modelos agregados, usa-se a regressão para analisar as variáveis, que podem assumir qualquer valor dentro da amplitude para variáveis contínuas. Quando os resultados dos comportamentos dos animais não são contínuos, os procedimentos de regressão padrão são inapropriados, uma vez que falta a linearidade e não se conhece a distribuição dos dados, sugere-se as análises de escolha qualitativa para esta situação.

As estimativas dos comportamentos individuais dos animais são mais precisas para conhecer assim suas preferências. A partir das observações individuais, que apresentam uma grande variação em cada fator e menos covariação entre os fatores do que as análises dos comportamentos dos $n$ grupos de animais através dos modelos agregados. A precisão é uma característica importante na estimação dos modelos, nos quais cada parâmetro pode ser estimado, geralmente com o incremento da variância da variável do modelo e, o decréscimo com a covariância entre as variáveis.

Uma aproximação geral para o estudo da seleção do recurso, proposta por McDonald et al. (1990), consiste em estimar a Função de Probabilidade da Seleção do Recurso (RSPF ), definida como a função que fornece as probabilidades de uso para unidades de recurso de diferentes tipos. Esta função pode ser estimada quando as alternativas dos recursos estudados são consideradas como um conjunto finito de unidades de recurso, mutuamente exclusivas, exaustivas e quando cada unidade pode ser caracterizada por um vetor de atributos.

Usando o modelo logit multinomial como um dos modelos de escolha discreta, para comportamentos humanos, pretende-se modificá-lo para as análises da seleção de recurso dos animais. Este modelo tem a vantagem de que a disponibilidade do recurso pode variar entre animais ao longo do tempo. Além disso, as características dos animais sob estudo podem ser incorporadas dentro do modelo. Freqüentemente na modelagem da seleção de recurso é assumido que um animal pode avaliar a qualidade de cada recurso disponível, sendo a qualidade uma função dos custos e benefícios associados com o recurso. Quando se determinam as preferências do recurso, é freqüentemente suposto que o animal selecionará um recurso de alta qualidade mais freqüentemente, que um de baixa qualidade.

Pretende-se aplicar as idéias de McFadden, (1974) (TRAIN, 2003), as quais fazem aplicações na economia para serem aplicadas na RSPF. Porém, na teoria econômica, as escolhas dos produtos de preferência não são muitas (TRAIN, 2003), apresentando em torno de 100 escolhas. Na situação dos animais, eles tem muitas escolhas e nem todos os recursos que a natureza oferece, são usados. Os grupos de animais escolhem a alternativa que fornece a máxima utilidade. O pesquisador não conhece todos os fatores relevantes

e também não conhece a função de utilidade exatamente. Quando os animais se agrupam, é duvidoso que os animais dentro de um grupo, selecionem independentemente outros membros do grupo para formar outro grupo. Neste caso, os grupos de animais poderiam ser considerados como os " *tomadores de decisões*". Se o tamanho do grupo é considerado associado com a Seleção do Recurso (RS), o tamanho do grupo pode ser um dos atributos medidos para os grupos observados (McCRACKEN et al., 1998).

Os animais ao selecionarem os alimentos ou habitat's utilizam o que a natureza oferece tendo múltiplas escolhas. Assim apresenta-se o modelo logit encaixado generalizado como uma proposta para que cada animal possa escolher e selecionar suas preferências de alimento ou habitat, devido a que o animal da mesma espécie têm diferentes preferências. Dessa forma, usando a maximização aleatória foi possível obter as estimativas das preferências dos animais, utilizando-se os dados das 28 corujas (*Strix occidentalis*), localizadas no noroeste de Califórnia, (USA). Sabe-se que uma floresta em climax é vital para a existência da coruja manchada, e essa espécie encontra-se em via de extinção (LAYMON et al., 1985; FORSMAN et al., 1984). Pretende-se selecionar as preferências de alimento e de habitat, pois cada coruja tem preferências que diferem das outras corujas. Generalizando esta idéia para os grupos de animais, poderemos conhecer suas preferências de habitat e/ou alimento, e assim evitar a extinção das espécies.

## 4.2 Desenvolvimento

### 4.2.1 Modelos de escolha qualitativa

Observa-se no Apêndice D1, (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados)) as situações de uso em ecologia do modelo de escolha qualitativa. Em qualquer situação de escolha, o animal faz a escolha usando dois ou mais items diferentes, ou alternativas que ele preferir. Na situação de escolha qualitativa, em que este modelo é usado para descrever as escolhas, sendo esse grupo definido como um grupo dentro dos grupos de animais que fazem a escolha, de um conjunto de alternativas, apresentando os seguintes critérios:

(i) O número de alternativas é um conjunto finito;

(ii) As alternativas são mutuamente exclusivas; sendo que, o animal escolhe uma alternativa dentro de um conjunto, o que implica que ele não escolherá outra alternativa;

(iii) O conjunto de alternativas é exaustivo, todas as possíveis alternativas são incluídas, e o animal, só escolhe necessariamente uma alternativa de um conjunto delas.

Quando a escolha não é descrita como qualitativa, as alternativas não são mutuamente exclusivas ou o conjunto de alternativas não é exaustivo. A expressão "situação de escolha discreta" é freqüentemente usada para indicar o mesmo que a "situação de escolha qualitativa", que resulta da distinção entre variáveis discretas e contínuas, denotando o conjunto de alternativas. Esta distinção também fornece o significado para o uso da expressão "qualitativa" na maioria das escolhas de interesse da população, que podem ser escolhidas em $b$ escolhas, uma depois da outra, ou em quanto tempo certo tipo de alimento seja selecionado. Os conjuntos de alternativas que são denotados pelas variáveis contínuas, consideram as escolhas de um

item, que não são quantitativas ou qualitativas, e o conjunto de alternativas pode ser denotado pelas variáveis discretas.

Modelos de escolha qualitativa são usados para analisar situações na qual os grupos dos animais podem observar as características da escolha entre um conjunto finito e exaustivo de alternativas mutuamente exclusivas. Além do mais, se as *b* escolhas dos grupos satisfazem esses critérios, então os modelos de escolha qualitativa são usados, somente se o número de alternativas é razoavelmente pequeno. Os modelos de escolha qualitativa que fazem parte dessa classe de modelos são o logit e o probit. Todos os modelos de escolha qualitativa calculam a probabilidade que os grupos de animais poderiam escolher uma alternativa particular de um conjunto de alternativas disponíveis, com os dados observados pelo pesquisador. Os modelos diferem na forma funcional que relaciona os dados observados à probabilidade. Denota-se os grupos de animais na situação de escolha qualitativa por *n* e o conjunto das alternativas vista como $J_n$. Este conjunto é algumas vezes chamado conjunto de escolha. É escrito como *n* para representar o fato que diferentes grupos de animais poderiam escolher diferentes conjuntos de alternativas em situações de escolhas similares.

Chamam-se as características observadas da alternativa *i*, como os grupos de animais que podem ser representados por um vetor $\mathbf{z}_{ni}$, $\forall i$ dentro de $J_n$, o conjunto de alternativas. Nota-se que as características de cada alternativa são subscritas por *n*, refletindo o fato de que diferentes grupos de animais escolhem alternativas com diferentes características. As escolhas dos grupos de animais, de uma alternativa óbvia, depende das características de cada alternativa disponível. Diferentes grupos de animais, de qualquer maneira, podem fazer diferentes escolhas usando as mesmas alternativas, desde que o valor relativo dos lugares que eles freqüentem apresentem características diferentes. As diferenças na avaliação de cada característica das alternativas dependem das características dos grupos de animais, observadas ou não pelo pesquisador. Chama-se o vetor das características observadas dos grupos de animais de $\mathbf{s}_n$.

A probabilidade de que grupos de animais escolham a alternativa *i* do conjunto $J_n$ (chamadas $P_{ni}$), depende das características observadas da alternativa *i* comparada com todas as outras alternativas (isto é $\mathbf{z}_{ni}$ relativa a todas $\mathbf{z}_{nj}$ para *j* dentro de $J_n$, $j \neq i$) e sobre $\mathbf{s}_n$. Modelos de escolha qualitativa especificam esta probabilidade como a função paramétrica de forma geral,

$$P_{ni} = f(\mathbf{z}_{ni}, \mathbf{z}_{nj}, \quad \forall j \text{ dentro de } J_n \text{ e } j \neq i, \ s_n, \ \boldsymbol{\beta}) \tag{4.1}$$

em que *f* é a função que relaciona os dados observados às probabilidades de escolha entre $\mathbf{z}_{ni}$ e $\mathbf{z}_{nj}$. Esta função é especificada sobre alguns vetores de parâmetros $\boldsymbol{\beta}$.

A descrição geral dos modelos de escolha qualitativa está completamente contida na expressão (4.1). Todos os modelos de escolha qualitativa têm essa forma geral, especificamente modelos de escolha qualitativa, tais como logit ou probit, que são obtidos pela especificação de *f*. A derivação dos modelos de escolha qualitativa da teoria de utilidade é baseada em uma preferência do comportamento dos grupos de animais e as análises do pesquisador. Considerando primeiro os grupos de animais, designando a utilidade da alternativa *i* dentro de $J_n$ como $U_{ni}$, e similarmente para cada outra alternativa dentro de $J_n$, esta utilidade depende de vários fatores, incluíndo as características da alternativa e as características dos grupos de animais. Definindo o vetor de todas as características relevantes da alternativa *i*, como vista pelo animal *a*, como $\mathbf{x}_{ai}$, e o vetor de todas as características relevantes do animal *a*, como $\mathbf{r}_a$, desde que $\mathbf{x}_{ai}$ e $\mathbf{r}_a$ incluam todos os fatores

relevantes, pode-se escrever a utilidade como função destes fatores,

$$U_{ni} = U(\mathbf{x}_{ai}, \mathbf{r}_a), \quad \forall i \text{ dentro de } J_n, \quad \text{a} = 1, \ldots, \text{A} \tag{4.2}$$

Thurstone (1927), originalmente desenvolveu os conceitos de estímulos psicológicos, levando a um modelo probit binário, de se os respondentes podem diferenciar os níveis do estímulo. Marschak (1960) interpretou os estímulos como uma utilidade e forneceu uma derivação a partir da maximização da utilidade. Segundo este autor, modelos que podem ser derivados desta maneira são chamados modelos de utilidade aleatória (RUMs). Note-se que modelos derivados da maximização da utilidade podem também ser usados para representar "tomada de decisões" que não envolva a maximização da utilidade. A derivação assegura que, o modelo é consistente com a maximização da utilidade, isto não impede que o modelo seja consistente com outras formas de comportamento. Os modelos também podem ser vistos como uma simples descrição da relação das variáveis explanatórias com o resultado de uma escolha, sem fazer referência de como exatamente a escolha foi feita.

### 4.2.2 O modelo logit encaixado (NL)

O modelo logit encaixado é apropriado quando o conjunto de alternativas pode ser particionado em subconjuntos, chamados encaixes, de tal maneira que as seguintes propriedades sejam satisfeitas:

- Para duas alternativas que estão no mesmo encaixe, a razão de probabilidades é independente dos atributos ou da existência de todas as outras alternativas, satisfazendo a propriedade *IIA* dentro de cada encaixe.

- Para duas alternativas em diferentes encaixes, a razão de probabilidades pode depender dos atributos de outras alternativas nos dois encaixes. Em geral a propriedade *IIA* não é satisfeita, para alternativas em diferentes encaixes.

Com dois encaixes, a propriedade *IIA* (Ver Apêndice D3 (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados))) satisfeita, dentro de cada encaixe mas, não através dos encaixes. O modelo logit encaixado com duas alternativas em um encaixe e duas alternativas alternadas dentro de outro encaixe, é apropriado para representar esta situação. A figura (4.1) apresenta o modelo logit encaixado de dois níveis, em que:

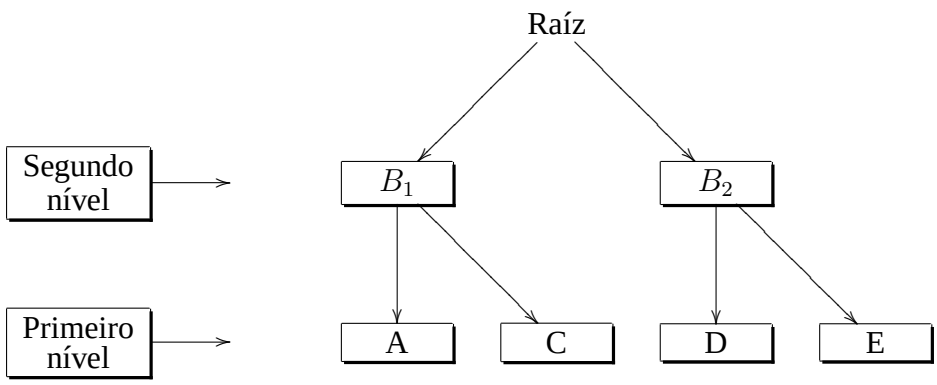



Figura 4.1 – Modelo logit encaixado de dois níveis

Daly e Zachary (1978), McFadden (1978) e Williams (1977) mostraram, independentemente e usando diferentes demonstrações, que o modelo logit encaixado é consistente com a maximização da utilidade. Seja o conjunto de alternativas *j*, particionado em *k* subconjuntos não sobrepostos chamados $B_1, \ldots, B_k$ denominados encaixes. A utilidade que o animal *n* escolha a alternativa *j* no encaixe $B_k$ é usualmente denotada, como $U_{nj} = V_{nj} + \varepsilon_{nj}$, em que $V_{nj}$ é a parte observada pelo pesquisador e $\varepsilon_{nj}$ é uma variável aleatória cujo valor não é observado pelo pesquisador (Ver Apêndice D2 (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados))). O modelo logit encaixado é obtido pela suposição de que o vetor da utilidade não observado, $\varepsilon_n = (\varepsilon_{n2}, \ldots, \varepsilon_{nJ})$, tem distribuição cumulativa,

$$exp\Big( - \sum_{k=1}^{K} \Big( \sum_{j \,\epsilon\, B_k} e^{\frac{-\varepsilon_{nj}}{\lambda_k}} \Big)^{\lambda_k} \Big) \tag{4.3}$$

em que $B_k$ é o encaixe das *j* alternativas particionadas nos *k* subconjuntos não sobrepostos.

O parâmetro $\lambda_k$ é a medida do grau de independência na utilidade não observada entre as alternativas do encaixe *k*.

Está distribuição é um tipo de distribuição de valor extremo generalizado (GEV). É uma generalização da distribuição que origina o modelo logit. No modelo logit, cada $\varepsilon_{nj}$ é independente com uma distribuição de valor extremo univariada. Já na GEV dada em (4.3), a distribuição marginal de cada $\varepsilon_{nj}$ é de valor extremo univariada. Entretanto, os $\varepsilon_{nj}$ são correlacionados dentro do encaixe. Para qualquer duas alternativas *j* e *g* no encaixe $B_k$, $\varepsilon_{nj}$ é correlacionado com $\varepsilon_{ng}$. Para qualquer duas alternativas em diferentes encaixes, a parte não observada da utilidade é ainda não correlacionada: Cov($\varepsilon_{nj}$, $\varepsilon_{ng}$) = 0, para qualquer $j \,\epsilon\, B_k$ e $g \,\epsilon\, B_l$ com $l \neq k$, e $k = 1, \ldots$, K, sendo *l* as alternativas do encaixe, e $l = 1, \ldots$, L.

O parâmetro $\lambda_k$ é uma medida do grau de independência na utilidade não observada entre as alternativas do encaixe *k*. Um valor alto de $\lambda_k$, significa maior independência e menos correlação. A estatística $1 - \lambda_k$ é uma medida de correlação.

McFadden (1978), afirma que $1 - \lambda_k$ pode ser usado como indicação de correlação. Um valor de $\lambda_k = 1$ indica completa independência dentro do encaixe *k*, isto é, não correlação. Quando $\lambda_k = 1, \ \forall k$, representando independência entre todas as alternativas dentro de todos os encaixes, a distribuição GEV torna-se o produto de termos de valor extremo independentes, cuja distribuição é dada por $F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}$. Neste caso o modelo logit encaixado se reduz ao modelo logit padrão. A distribuição para as componentes não observadas da utilidade, origina a seguinte probabilidade de escolha para a alternativa $i \,\epsilon\, B_k$:

$$P_{ni} = \frac{e^{\frac{V_{ni}}{\lambda_k}} \Big( \sum_{j \,\epsilon\, B_k} e^{V_{nj}/\lambda_k} \Big)^{\lambda_k - 1}}{\sum_{l=1}^{L} \Big( \sum_{j \,\epsilon\, B_l} e^{\frac{V_{nj}}{\lambda_l}} \Big)^{\lambda_l}} \tag{4.4}$$

Pode-se usar esta fórmula para mostrar que a propriedade *IIA* se mantém dentro de cada subconjunto de alternativas, mas não através dos subconjuntos. Considere as alternativas $i \,\epsilon\, B_k$ e $g \,\epsilon\, B_l$, como o denominador de (4.4) é o mesmo para todas as alternativas, a razão de probabilidades é a razão dos numera-

dores:

$$\frac{P_{ni}}{P_{ng}} = \frac{e^{\frac{V_{ni}}{\boldsymbol{\lambda}_k}} \left( \sum_{j \,\epsilon\, B_k} e^{\frac{V_{nj}}{\boldsymbol{\lambda}_k}} \right)^{\boldsymbol{\lambda}_k - 1}}{e^{\frac{V_{ng}}{\boldsymbol{\lambda}_l}} \left( \sum_{j \,\epsilon\, B_l} e^{\frac{V_{nj}}{\boldsymbol{\lambda}_l}} \right)^{\boldsymbol{\lambda}_l - 1}}.$$

Se $k = l$, isto é, $i$ e $g$ estão no mesmo encaixe, então:

$$\frac{P_{ni}}{P_{ng}} = \frac{e^{\frac{V_{ni}}{\boldsymbol{\lambda}_k}}}{e^{\frac{V_{ng}}{\boldsymbol{\lambda}_l}}}.$$

Está razão é independente de todas as outras alternativas. A razão de probabilidades depende dos atributos de todas as alternativas nos encaixes que contém $i$ e $g$. Quando $\boldsymbol{\lambda}_k = 1, \forall\, k, (1 - \boldsymbol{\lambda}_k = 0)$, indicando não correlação entre as componentes não observadas de utilidade para as alternativas dentro do encaixe, as probabilidades de escolha tornam-se um logit simples. O modelo logit encaixado é uma generalização do logit que leva em consideração um padrão particular de correlação na utilidade não observada.

O parâmetro $\boldsymbol{\lambda}_k$ pode diferir para diferentes encaixes, refletindo diferentes correlações entre fatores não observados dentro de cada encaixe. O valor de $\boldsymbol{\lambda}_k$, deve estar dentro de uma amplitude particular para o modelo ser consistente com o comportamento de maximização da utilidade. Se $\boldsymbol{\lambda}_k$, $\forall\, k$ está entre zero e um, o modelo é consistente com a maximização da utilidade para todos os possíveis valores das variáveis explanatórias. Para $\boldsymbol{\lambda}_k > 1$, o modelo é consistente com o comportamento de maximização da utilidade para algumas amplitudes das variáveis explanatórias, mas não para todos os valores. Kling e Herriges (1995), Herriges e Kling (1996) apresentam testes de consistência do modelo logit encaixado com a maximização da utilidade quando $\boldsymbol{\lambda}_k > 1$. Um valor negativo de $\boldsymbol{\lambda}_k$ é inconsistente com a maximização da utilidade e implica que o melhoramento dos atributos de uma alternativa pode decrescer a probabilidade da alternativa ser escolhida. Com $\boldsymbol{\lambda}_k$ positivo, o logit encaixado aproxima, o modelo de Tversky (1972) quando, $\boldsymbol{\lambda}_k \rightarrow 0$.

### 4.2.3 O modelo de valor extremo heterocedástico

Os modelos de escolha de modo de viagem interurbano são baseados na hipótese de maximização da utilidade a qual assume que a escolha de modo de viagem de um indivíduo reflete as preferências subjacentes para cada uma das alternativas disponíveis, e que o indivíduo seleciona a alternativa com a mais alta preferência ou utilidade. A utilidade que um indivíduo associa com uma alternativa é especificada pela soma de uma componente determinística (dependendo dos atributos observados da alternativa e do indivíduo) e uma componente aleatória (que representa os efeitos dos atributos não observados do indivíduo e as características não observadas da alternativa) (BHAT, 1995). Na maioria dos modelos de escolha de modo interurbano, as componentes aleatórias das utilidades das diferentes alternativas são assumidas por ser independentes e identicamente distribuídas ($IID$) com uma distribuição de valor extremo tipo I (JOHNSON; KOTZ, 1970).

O modelo valor extremo heterocedástico permite considerar variâncias diferentes nas componentes aleatórias. As alternativas variâncias desiguais nas componentes aleatórias geralmente ocorrem quando a variância de uma variável não observada que afeta a escolha é diferente para alternativas diferentes. Daganzo (1979), usou independentemente a distribuição exponencial negativa com variâncias e alternativas diferentes

para as componentes de erro aleatórios nos modelos de escolha discreta de forma fechada os quais não garantem a propriedade $IIA$ ( Anexo A3, http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados)). A utilidade aleatória da alternativa i, $U_i$, para um indivíduo nos modelos de utilidade aleatória, apresentam a forma,

$$U_i = V_i + \boldsymbol{\epsilon}_i \tag{4.5}$$

em que $V_i$ é a componente sistemática da utilidade da alternativa *i*, a qual é função dos atributos observados da alternativa *i* e das características observadas do indivíduo, $\boldsymbol{\epsilon}_i$ é a componente aleatória da função de utilidade. Seja $C$ o conjunto de alternativas disponíveis ao indivíduo. Assumimos que as componentes aleatórias nas utilidades das diferentes alternativas têm distribuição de valor extremo tipo I e são independentes, porém distribuídas não identicamente. Nós também assumimos que as componentes aleatórias têm parâmetro de localização igual a zero e parâmetro de escala igual a $\boldsymbol{\theta}_i$ para a *i*-ésima alternativa e,

$$f(\boldsymbol{\epsilon}_i) = \frac{1}{\boldsymbol{\theta}_i} e^{-\frac{\boldsymbol{\epsilon}_i}{\boldsymbol{\theta}_i}} e^{-e^{-\frac{\boldsymbol{\epsilon}_i}{\boldsymbol{\theta}_i}}}, \qquad e \qquad F_i(z) = \int_{\boldsymbol{\epsilon}_i=-\infty}^{\boldsymbol{\epsilon}_i=z} f(\boldsymbol{\epsilon}_i) d\boldsymbol{\epsilon}_i = e^{-e^{-\frac{z}{\boldsymbol{\theta}_i}}} \tag{4.6}$$

A formulação da utilidade aleatória da equação (4.5) combinada com a distribuição de probabilidade assumida para as componentes aleatórias em (4.6), e com dependência assumida entre as componentes aleatórias das diferentes alternativas, nos permitem desenvolver a probabilidade de que um indivíduo escolherá a alternativa *i*, $P_i$ do conjunto $C$ de alternativas disponíveis,

$$\begin{aligned} P_i &= \text{Prob}(U_j > U_i), \ \text{ para todo } \ j \neq i \ \varepsilon \ C \\ &= \text{Prob}(\boldsymbol{\epsilon}_j \leq V_i - V_j + \boldsymbol{\epsilon}_i), \ \text{ para todo } \ j \neq i \ \varepsilon \ C \\ &= \int_{\boldsymbol{\epsilon}_i=-\infty}^{\boldsymbol{\epsilon}_i=+\infty} \prod_{j \ \varepsilon \ C \ (j \neq i)} \Lambda\left[\frac{V_i - V_j + \boldsymbol{\epsilon}_i}{\boldsymbol{\theta}_j}\right] \frac{1}{\boldsymbol{\theta}_i} \lambda\left(\frac{\boldsymbol{\epsilon}_i}{\boldsymbol{\theta}_i}\right) d\boldsymbol{\epsilon}_i, \end{aligned} \tag{4.7}$$

Em que $\lambda[.]$ e $\Lambda[.]$ são respectivamente a função de densidade de probabilidade e a função de distribuição cumulativa da distribuição do valor extremo tipo I padrão, em que:

$$\lambda(t) = e^{-t} e^{-e^{-t}}, \quad \Lambda(t) = e^{-e^{-t}}. \tag{4.8}$$

Substituindo $w = \frac{\boldsymbol{\epsilon}_i}{\boldsymbol{\theta}_i}$, na equação (4.8), a probabilidade de escolher a alternativa $i$ pode ser reescrita como,

$$P_i = \int_{w=-\infty}^{w=+\infty} \prod_{j \ \varepsilon \ C \ (j \neq i)} \Lambda\left[\frac{V_i - V_j + \boldsymbol{\theta}_i w}{\boldsymbol{\theta}_j}\right] \lambda(w) dw. \tag{4.9}$$

As probabilidades dadas pela expressão (4.9) somam 1 sobre todas as alternativas. Apresentando que as probabilidades de escolha das alternativas somam 1 no modelo do valor extremo heterocesdastico, e fazendo un exemplo para três alternativas, mas isto pode ser generalizado para diferente número de alternativas. Para três casos de alternativas, podemos escrever as probabilidades de escolha baseadas na expressão (4.9), como,

$$\begin{aligned} P_1 &= \int_{z=-\infty}^{z=+\infty} \Lambda\Big[\frac{V_1 - V_2 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_2}\Big] \Lambda\Big[\frac{V_1 - V_3 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_3}\Big] \lambda(z) dz \\ P_2 &= \int_{z=-\infty}^{z=+\infty} \Lambda\Big[\frac{V_2 - V_1 + \boldsymbol{\theta}_2 z}{\boldsymbol{\theta}_1}\Big] \Lambda\Big[\frac{V_2 - V_3 + \boldsymbol{\theta}_2 z}{\boldsymbol{\theta}_3}\Big] \lambda(z) dz \\ P_3 &= \int_{z=-\infty}^{z=+\infty} \Lambda\Big[\frac{V_2 - V_1 + \boldsymbol{\theta}_2 z}{\boldsymbol{\theta}_1}\Big] \Lambda\Big[\frac{V_2 - V_3 + \boldsymbol{\theta}_2 z}{\boldsymbol{\theta}_3}\Big] \lambda(z) dz. \end{aligned} \tag{4.10}$$

Definindo a função de H($z$), assim,

$$H(z) \quad = \quad \Lambda(z)\Lambda\Big[\frac{V_1 - V_2 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_2}\Big]\Lambda\Big[\frac{V_1 - V_3 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_3}\Big]. \tag{4.11}$$

H($z$) é o produto da funções de distribuição acumuladas própias (product of proper cumulative distributions functions), isto é também, a função de distribuição acumulada própia, mas pode ser escrita assim,

$$\int_{z=-\infty}^{z=+\infty} \frac{d}{dz} H(z)dz = H(+\infty) - H(-\infty) = 1. \tag{4.12}$$

fazendo a diferenciação de H($z$) com respeito a $z$ e usando a expressão 4.13, podemos escrever,

$$\begin{aligned}\int_{z=-\infty}^{z=+\infty} \frac{d}{dz} H(z)dz = \quad & \textstyle\int_{z=-\infty}^{z=+\infty} \quad \lambda(z)\Lambda\left[\dfrac{V_1 - V_2 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_2}\right]\Lambda\left[\dfrac{V_1 - V_3 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_3}\right]dz + \\ & \textstyle\int_{z=-\infty}^{z=+\infty} \quad \Lambda(z)\lambda\left[\dfrac{V_1 - V_2 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_2}\right]\Lambda\left[\dfrac{V_1 - V_3 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_3}\right]\left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_2}dz\right) + \\ & \textstyle\int_{z=-\infty}^{z=+\infty} \quad \Lambda(z)\lambda\left[\dfrac{V_1 - V_2 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_2}\right]\Lambda\left[\dfrac{V_1 - V_3 + \boldsymbol{\theta}_1 z}{\boldsymbol{\theta}_3}\right]\left(\dfrac{\boldsymbol{\theta}_1}{\boldsymbol{\theta}_3}dz\right).\end{aligned} \tag{4.13}$$

Fazendo mudanças nas variáveis $z'' = \frac{(V_1 - V_3 + \boldsymbol{\theta}_1 z)}{\boldsymbol{\theta}_3}$, no termo da segunda integral e $z' = \frac{(V_1 - V_2 + \boldsymbol{\theta}_1 z)}{\boldsymbol{\theta}_2}$, no termo da terceira integral. O primeiro termo da expressão (4.13) é $P_1$, o segundo é $P_2$ e o terceiro é $P_3$, isto é, $P_1 + P_2 + P_3 = 1$.

Se os parâmetros de escala das componentes aleatórias de todas as alternativas são iguais, então a probabilidade da expressão (4.9) iguala aquela do modelo logit multinomial (McFADDEN, 1974). Formalmente, o efeito e uma pequena alteração na utilidade sistemática da alternativa $l$ sobre a probabilidade de escolher a alternativa $i$, pode ser escrita como,

$$\frac{\partial P_i}{\partial V_i} \quad = \quad \int_{w=-\infty}^{w=+\infty} -\frac{1}{\boldsymbol{\theta}_l}\exp\left[\frac{V_i - V_j + \boldsymbol{\theta}_i w}{\boldsymbol{\theta}_l}\right] \prod_{j\ \varepsilon\ C\ (j \neq i)} \Lambda\left[\frac{V_i - V_j + \boldsymbol{\theta}_i w}{\boldsymbol{\theta}_j}\right]\lambda(w)dw \tag{4.14}$$

e o efeito de uma variação na utilidade sistemática da alternativa $i$, sobre a probabilidade de escolher $i$ pode ser escrita como,

$$\frac{\partial P_i}{\partial V_i} = -\sum_{l\ \varepsilon\ C,\ l \neq i} \frac{\partial P_i}{\partial V_i}. \tag{4.15}$$

Assumindo linearidade nos parâmetros, para a componente sistemática da utilidade de todas as alternativas, a elasticidade cruzada para a alternativa $i$ com respeito a uma alteração no $k$-ésimo nível da variável na $l$-ésima utilidade sistemática, das alternativas $x_{kl}$, pode ser obtida como,

$$\eta_{xkl}^{P_i} = \left[\frac{\partial P_i}{\partial V_l}/P_i\right] \times \boldsymbol{\beta}_k \times x_{kl} \tag{4.16}$$

em que $\boldsymbol{\beta}_k$ é o coeficiente estimado no nível da variável *k*. A correspondente auto-elasticidade para a alternativa *i* com respeito a uma alteração em $x_{ki}$ é,

$$\eta_{xki}^{P_i} = \left[\frac{\partial P_i}{\partial V_l}/P_i\right] \times \boldsymbol{\beta}_k \times x_{ki}. \tag{4.17}$$

O modelo de valor extremo heterocedástico desenvolvido por Bhat (1995), é estimado usando a técnica de máxima verossimilhança. Assume-se uma especificação linear nos parâmetros para a utilidade sistemática de cada alternativa dada por, $V_{qi} = \boldsymbol{\beta}' X_{qi}$, para o *q*-ésimo indivíduo e a *i*-ésima alternativa. Os parâmetros a serem estimados no modelo heterocedástico são o vetor de parâmetros $\boldsymbol{\beta}$ e os parâmetros de escala da componente aleatória de cada uma das alternativas. A função do logaritmo da verossimilhança a ser maximizada pode ser escrita como,

$$L = \sum_{q=1}^{q=Q} \sum_{i \in C_q} y_{qi} \log \left\{ \int_{w=-\infty}^{w=+\infty} \prod_{j \in C_q,\, j \neq i} \Lambda \left[\frac{V_{qi} - V_{qj} + \theta_i w}{\theta_j}\right] \lambda(w) dw \right\} \tag{4.18}$$

em que $C_q$ é o conjunto de escolha das alternativas disponíveis para o *q*-ésimo indivíduo e $y_{qi}$, é definido como,

$$y_{qi} = \begin{cases} 1 & \text{se o } q\text{-ésimo indivíduo escolhe a alternativa } i \\ & q=1,2,\ldots,Q,\ i=1,2,\ldots,I \\ 0 & \text{de outra maneira,} \end{cases}$$

A função logaritmo da verossimilhança da expressão (4.18), não tem forma fechada. Uma integral imprópria precisa ser calculada para cada combinação alternativa-indivíduo em cada iteração, da maximização da função logaritmo da verossimilhança. O uso de técnicas da integração numérica convencionais, tais como o método de Simpson ou a integração de Romberg, para avaliar tais integrais é complicada, custosa e freqüentemente leva a estimativas instáveis, devido a que elas requerem a avaliação do integrando em um grande número de intervalos espaçados igualmente na reta real (BUTLER, MOFFITT, 1982; CHINTAGUNTA et al., 1991).

Por outro lado a quadratura Gaussiana (PRESS et al., 1986) é um procedimento mais sofisticado que pode obter estimativas altamente precisas das integrais na função de verossimilhança, avaliando o integrando em um número relativamente pequeno de pontos de suporte. Para aplicar o método da quadratura Gaussiana precisamos definir a variável, $u$=$e^{-w}$, então $\lambda(w)dw = -e^u du$, e $w = -ln(u)$. Definindo a função $G_{qi}$, como,

$$G_{qi}(u) = \prod_{j \in C_q,\, j \neq i} \lambda \left[\frac{V_{qi} - V_{qj} - \boldsymbol{\theta}_i ln(u)}{\boldsymbol{\theta}_j}\right] \tag{4.19}$$

a expressão (4.18), pode ser escrita como,

$$L = \sum_{q} \sum_{i \in C_q} y_{qi} \log \left\{ \int_{u=0}^{u=\infty} G_{qi}(u) e^{-u} du \right\}. \tag{4.20}$$

A expressão entre parênteses na equação (4.20) pode ser estimada usando a fórmula da quadratura Gaussiana de Laguerre, a qual substitui a integral por uma soma de termos sobre certo número ($K$) de pontos

suporte ( PRESS et al., 1986), ( STROUD; SECREST, 1966). Quando as componentes estocásticas de utilidade são heterocedásticas e independentes, usamos o Modelo de Valor Extremo Heterocedástico (HEV). O modelo de HEV assume que a utilidade da alternativa *j* para cada indivíduo *i* tem componentes aleatórias heterocedásticas.

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

a distribuição acumulada da distribuição Gumbel $\boldsymbol{\epsilon}_{ij}$ é,

$$F(\boldsymbol{\epsilon}_{ij}) = \exp(-\exp(\boldsymbol{\epsilon}_{ij}/\boldsymbol{\theta}_j))$$

A variância de $\boldsymbol{\epsilon}_{ij}$ é $\frac{1}{6}\pi^2\boldsymbol{\theta}_j^2$. O erro da variância é proporcional ao quadrado do parâmetro de escala e ao menos um dos parâmetros de escala poderia ser normalizado para 1. O Software SAS estima o modelo HEV sob a restrição de escala reduzida para o modo "1", ($\boldsymbol{\theta}_1 = 1$).

Bhat (1995), apresenta as componentes aleatórias da função utilidade, a ser não idênticas. Especificamente, o modelo HEV assume independência, porém com os termos do erro não idênticos, e com a distribuição do valor extremo tipo I. O modelo HEV apresenta as variâncias das componentes aleatórias de utilidade, para diferir das alternativas cruzadas. Ele argumenta que o modelo não tem a propriedade $IIA$, e este modelo contém o modelo condicional como um caso especial (BHAT, 1995).

Primeiro usa-se o método de Newton-Raphson (Ver Apêndice C3-1). O segundo método de integração da quadratura gaussiana, Gauss-Laguerre $INTEGRATE = LAGUERRE$ especifica a restrição de escala reduzida e variância única para todas as escolhas (Ver Apêndice C3-2). O terceiro método de integração adaptativa usado aqui é o Hardy $INTEGRATE = HARDY$, que usa o tipo-Romberg adaptive. Este procedimento de integração adaptativa produz mais exatidão nos cálculos da probabilidade e os valores da função de verossimilhança, mas não é muito utilizado pois requer um tempo considerável de simulação (Ver Apêndice C3-3).

Registra-se que estes procedimentos encontram-se implementados no sistema computacional SAS, através do procedimento MDC, no qual identifica-se as variáveis que contém as possíveis escolhas para cada coruja. A opção do HEV, requer especificação do $intorder$ que específica os números dos termos da soma para a integração da quadratura gaussiana, o máximo da ordem é 45, esta opção aplica-se somente ao método de integração Laguerre.

O procedimento MDC apresenta o resumo da especificação do modelo, o método de otimização usado é o Dual Quasi-Newton que usa o fator de Cholesky para aproximar a matriz Hessiana, (o algoritmo original de Quasi-Newton usa a aproximação da inversa Hessiana), este método de otimização usa o gradiente e portanto não é necessário fazer as derivadas de segunda ordem desde que este seja aproximado.

Nós três procedimentos de estimação, usa-se a opção $PRED$ que refere-se as probabilidades de decisão dos métodos empregados, e apresenta-se em gráficos no Anexo D (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados)), nos quais são feitas escolha por escolha para cada coruja, também em cada gráfico observa-se o valor da mediana das probabilidades de decisão de cada escolha, para os três métodos de simulação implementados.

O procedimento MDC apresenta nove medidas de qualidade de ajuste para o modelo de escolha discreta. Sete medidas destas são pseudo-$R^2$ baseadas na hipótese nula de que todos os coeficientes exceto que

para um termo do intercepto são nulos. Estrella (1998), propõe o uso desta medida de qualidade de ajuste análoga ao $R^2$, para uso quando as variáveis dependentes são dicotômicas. O índice da razão da verossimilhança McFadden's ($LRI$) é o menor valor. Mckelvey e Zavoina (1975), Aldrich e Nelson (1984) e Dhrymes (1996), propõem o uso de medidas de pseudo-$R^2$, os coeficientes do logit podem ser usados para calcular a variância explicada dos valores previstos para a variável dependente latente, denotada por $var(\hat{y}_i)$, em que a perturbação tem variância única. A variância total reduz a variância explicada mais um, e a pseudo-$R^2$ de Mckelvey e Zavoina torna-se,

$$R^2 = \frac{var(\hat{y}_i)}{1 + var(\hat{y}_i)}$$

Aldrich-Nelson (1984) e Dhrymes (1996) usam a pseudo-$R^2$, e baseada na razão de verossimilhança, eles utilizam o valor da verossimilhança para o modelo nulo, denotada como $L_0$, depois acham o valor da verossimilhança para o modelo completo, denotada como $L_1$. A medida de Aldrich-Nelson usa a estatística $\chi^2$, 2LLR definida como $-2log(L_0/L_1)$, com *k* o número de variáveis independentes estimadas, graus de liberdade e $N$ sendo o número de observações, a pseudo-$R^2$ de Aldrich-Nelson, definida como,

$$R^2 = \frac{-2log(L_0/L_1)}{N - 2log(L_0/L_1)} = \frac{-2LLR}{N - 2LLR}$$

tem distribuição $\chi^2$ e a pseudo-$R^2$ são baseadas nas medidas de verossimilhança que são comparadas quando é incluído o termo constante. A estatística Estrella (ajustada usa o Critério de Informação de Akaike (AIC)) e também baseada na verossimilhança, mas a interpretação é similar à estatística $R^2$, usada no contexto de regressão linear.

O índice de razão de verossimilhança é definido como,

$$\bar{p}^2 = 1 - \frac{L(M) - K}{L(C)}$$

em que $L(M)$ é modelo do valor do logaritmo da verossimilhança, $L(C)$ é o valor do logaritmo da verossimilhança com apenas as constantes das alternativas específicas e a matriz de covariância, (BHAT, 1995)

### 4.2.4 Análises dos dados da coruja manchada usando o modelo HEV

Para ilustrar a aplicação dos métodos descritos usa-se um conjunto de dados na área de ecologia. Os dados desse exemplo referem-se a coruja manchada ,(*Strix occidentalis*) na figura ((2.1)(a), capítulo 2), e a figura ((2.1)(b)), da propriedade da companhia Green Diamond Resource Company ($GDRCo$), localizada nos condados de Del Norte e Humboldt, no noroeste da California, USA. Observa-se na figura ((2.1)(c)) todas as localizações geográficas de todas as 28 corujas manchadas. Observa-se na tabela ((2.1), capítulo 2), as variáveis explanatórias das corujas manchadas, as variáveis "acl" foram incluídas centrando um buffer circular, em um ponto em que o raio é de 250 m.

Na tabela (4.1), as corujas manchadas 13 e 9 movimentaram-se mais, e as 15 e 28 se movimentaram menos. Observa-se no Anexo A.1.1 (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados)), as localizações usando as mesmas coordenadas dos eixos X e Y, para assim conhecer o seu posicionamento. As corujas 1, 2; 3, 4, 5; 6, 7, 8, 9, 10, 11; 17, 18, 21, 22, 25, 26 e 23, 24, 28 ficaram em lugares

Tabela 4.1 – Número de movimentações (n), e número de escolhas de cada coruja manchada

| Corujas | n | Escolhas | Corujas | n | Escolhas |
|---|---|---|---|---|---|
| 1 | 457 | 8 | 15 | 59 | 4 |
| 2 | 716 | 7 | 16 | 150 | 7 |
| 3 | 284 | 6 | 17 | 228 | 9 |
| 4 | 141 | 4 | 18 | 439 | 8 |
| 5 | 143 | 3 | 19 | 350 | 7 |
| 6 | 319 | 7 | 20 | 74 | 7 |
| 7 | 455 | 3 | 21 | 314 | 9 |
| 8 | 423 | 7 | 22 | 186 | 7 |
| 9 | 746 | 7 | 23 | 281 | 8 |
| 10 | 109 | 3 | 24 | 204 | 8 |
| 11 | 281 | 8 | 25 | 154 | 8 |
| 12 | 253 | 7 | 26 | 100 | 8 |
| 13 | 1226 | 6 | 27 | 13 | 3 |
| 14 | 350 | 6 | 28 | 97 | 8 |

similares. Neste anexo, podemos observar que as corujas manchadas, sendo da mesma espécie, e estando na mesma área, movimentaram-se diferentes vezes, como aparece em cada gráfico. O $n$ refere-se aos números de movimentações de cada coruja manchada, o que nos faz supor que, cada coruja manchada tem escolhas independentes. As análises das movimentações individuais de cada coruja manchada com seus respectivos gráficos, e as análises de cada uma delas encontram-se disponíveis (http://www.lce.esalq.usp.br/tadeu.html (Teses e Dissertações de orientados)).

### 4.2.5 Discussão

A Função de seleção de recurso aleatória para várias escolhas das corujas que selecionaram diferentes escolhas entre elas e independentes, usou-se o modelo de escolha discreta (DCM), do sistema computacional SAS, no qual se comparam diferentes métodos de estimação, como pode-se observar nas tabelas (4.2, 4.3, 4.4, 4.5). Os métodos de estimação usados, são, o Método de Newton-Raphson, método de integração da quadratura gaussiana, Gauss-Laguerre e método de integração adaptativa de Hardy. Por exemplo da tabela (4.2), a coruja manchada 1, observa-se que ao ajustar o modelo usando o método Newton-Raphson ele apresenta como a melhor probabilidade de decisão a escolha 3, que refere-se á variável acl_p2 (Proporção da zona tampão de bosque de idade entre 6 e 20 anos), tabela (4.2), e usando os métodos de quadratura Gaussiana Gauss-Laguerre e Hardy, a melhor probabilidade de decisão corresponde á escolha 7 que refere-se a acl_pscv (Porcentagem do coeficiente de variação do tamanho da mancha), neste exemplo o melhor Index corresponde ao método de integração adaptativa de Hardy, por ser maior com respeito aos outros Index.

Dos 16 conjuntos de dados selecionados das corujas manchadas, 8 apresentaram os melhores Index do Método de Hardy, 6 usando o Método Newton-Raphson, e 2 o método Laguerre. Observa-se que as variáveis das corujas manchadas selecionadas, usando os métodos de simulação descritos nas tabelas (4.2, 4.3, 4.4, 4.5) que pelo menos selecionaram como mínimo 2 preferências pelas variáveis. A tabela (4.6), apresentam as variáveis selecionadas usando os métodos de simulação anteriores.

Tabela 4.2 – Comparação dos métodos de simulação das corujas manchadas

| | | | | Probabilidades de decisão | | |
|---|---|---|---|---|---|---|
| Coruja | Variáveis | n | Escolhas | Newton-Raphson | Gauss-Laguerre | Hardy |
| 1 | drd_cent | 23×8 | 1 | 0,105 | 0,114 | 0,110 |
| | acl_p4 | 54×8 | 2 | 0,107 | 0,113 | 0,112 |
| | acl_p2 | 81×8 | 3 | 0,116 | 0,110 | 0,109 |
| | acl_p3 | 66×8 | 4 | 0,092 | 0,111 | 0,109 |
| | rwd_cent | 16×8 | 5 | 0,098 | 0,113 | 0,112 |
| | wwb_cent | 155×8 | 6 | 0,101 | 0,109 | 0,108 |
| | acl_pscv | 18×8 | 7 | 0,101 | 0,115 | 0,114 |
| | slp_cent | 44×8 | 8 | 0,096 | 0,110 | 0,109 |
| | | 457×8 | | | | |
| | Index | | | 0,162 | 0,166 | 0,169 |
| 2 | drd_cent | 51×7 | 1 | 0,105 | 0,141 | 0,143 |
| | rwd_cent | 69×7 | 2 | 0,110 | 0,113 | 0,111 |
| | acl_p3 | 143×7 | 3 | 0,133 | 0,119 | 0,118 |
| | acl_p2 | 43×7 | 4 | 0,088 | 0,104 | 0,104 |
| | acl_p4 | 62×7 | 5 | 0,137 | 0,123 | 0,122 |
| | slp_cent | 144×7 | 6 | 0,092 | 0,115 | 0,113 |
| | hgt_cent | 204×7 | 7 | 0,101 | 0,110 | 0,108 |
| | | 716×7 | | | | |
| | Index | | | 0,281 | 0,238 | 0,240 |
| 3 | slp_cent | 5×6 | 1 | 0,154 | 0,180 | 0,182 |
| | age_cent | 45×6 | 2 | 0,165 | 0,190 | 0,187 |
| | acl_p4 | 52×6 | 3 | 0,173 | 0,188 | 0,193 |
| | acl_pscv | 33×6 | 4 | 0,150 | 0,169 | 0,169 |
| | drd_cent | 82×6 | 5 | 0,130 | 0,141 | 0,158 |
| | acl_p2 | 67×6 | 6 | 0,067 | 0,140 | 0,157 |
| | | 284×6 | | | | |
| | Index | | | 0,266 | 0,176 | 0,182 |
| 8 | acl_p2 | 46×7 | 1 | 0,154 | 0,133 | 0,130 |
| | age_cent | 70×7 | 2 | 0,136 | 0,122 | 0,122 |
| | acl_p4 | 52×7 | 3 | 0,136 | 0,124 | 0,123 |
| | slp_cent | 93×7 | 4 | 0,133 | 0,120 | 0,119 |
| | acl_ed | 90×7 | 5 | 0,130 | 0,122 | 0,121 |
| | wwb_cent | 32×7 | 6 | 0,142 | 0,122 | 0,121 |
| | acl_p3 | 40×7 | 7 | 0,113 | 0,116 | 0,115 |
| | | 423×7 | | | | |
| | Index | | | 0,058 | 0,091 | 0,005 |
| 9 | acl_p2 | 73×7 | 1 | 0,124 | 0,101 | 0,097 |
| | slp_cent | 53×7 | 2 | 0,101 | 0,104 | 0,101 |
| | res_cent | 42×7 | 3 | 0,088 | 0,111 | 0,107 |
| | phw_cent | 101×7 | 4 | 0,122 | 0,101 | 0,098 |
| | rwd_cent | 72×7 | 5 | 0,126 | 0,099 | 0,088 |
| | pww_cent | 81×7 | 6 | 0,124 | 0,102 | 0,098 |
| | drd_cent | 324×7 | 7 | 0,120 | 0,102 | 0,098 |
| | | 746×7 | | | | |
| | Index | | | 0,250 | 0,300 | 0,303 |

Tabela 4.3 – Comparação dos métodos de simulação das corujas manchadas

| | | | | Probabilidades de decisão | | |
|---|---|---|---|---|---|---|
| Coruja | Variáveis | n | Escolhas | Newton-Raphson | Gauss-Laguerre | Hardy |
| 11 | acl_ed | 19×8 | 1 | 0,118 | 0,083 | 0,085 |
| | drd_cent | 28×8 | 2 | 0,170 | 0,077 | 0,126 |
| | acl_mps | 4×8 | 3 | 0,171 | 0,082 | 0,084 |
| | prw_cent | 22×8 | 4 | 0,126 | 0,085 | 0,087 |
| | age_cent | 87×8 | 5 | 0,065 | 0,073 | 0,074 |
| | acl_pscv | 71×8 | 6 | 0,123 | 0,083 | 0,085 |
| | acl_p2 | 29×8 | 7 | 0,128 | 0,081 | 0,082 |
| | acl_p4 | 21×8 | 8 | 0,100 | 0,077 | 0,078 |
| | | 281×8 | | | | |
| | Index | | | 0,216 | 0,211 | 0,213 |
| 12 | acl_mps | 16×6 | 1 | 0,168 | 0,128 | 0,127 |
| | slp_cent | 27×6 | 2 | 0,170 | 0,128 | 0,126 |
| | drd_cent | 23×6 | 3 | 0,171 | 0,132 | 0,130 |
| | acl_dne | 53×6 | 4 | 0,171 | 0,133 | 0,132 |
| | acl_ed | 35×6 | 5 | 0,174 | 0,135 | 0,134 |
| | acl_p4 | 99×6 | 6 | 0,173 | 0,135 | 0,134 |
| | | 253×6 | | | | |
| | Index | | | 0,200 | 0,228 | 0,233 |
| 13 | age_cent | 55×6 | 1 | 0,179 | 0,100 | 0,116 |
| | rwd_cent | 128×6 | 2 | 0,219 | 0,099 | 0,118 |
| | wwb_cent | 80×6 | 3 | 0,214 | 0,099 | 0,118 |
| | acl_pscv | 154×6 | 4 | 0,225 | 0,115 | 0,115 |
| | drd_cent | 208×6 | 5 | 0,218 | 0,098 | 0,117 |
| | hgt_cent | 601×6 | 6 | 0,222 | 0,099 | 0,116 |
| | | 1226×6 | | | | |
| | Index | | | 0,200 | 0,316 | 0,314 |
| 14 | hgt_cent | 32×6 | 1 | 0,160 | 0,155 | 0,155 |
| | drd_cent | 57×6 | 2 | 0,152 | 0,161 | 0,160 |
| | acl_pscv | 66×6 | 3 | 0,161 | 0,167 | 0,166 |
| | acl_dne | 36×6 | 4 | 0,183 | 0,166 | 0,166 |
| | acl_p4 | 63×6 | 5 | 0,173 | 0,166 | 0,166 |
| | slp_cent | 96×6 | 6 | 0,176 | 0,168 | 0,167 |
| | | 350×6 | | | | |
| | Index | | | 0,197 | 0,190 | 0,191 |
| 16 | acl_p2 | 19×7 | 1 | 0,130 | 0,124 | 0,123 |
| | slp_cent | 30×7 | 2 | 0,084 | 0,123 | 0,122 |
| | acl_pscv | 14×7 | 3 | 0,118 | 0,137 | 0,137 |
| | acl_p4 | 20×7 | 4 | 0,111 | 0,129 | 0,129 |
| | hgt_cent | 15×7 | 5 | 0,139 | 0,114 | 0,113 |
| | age_cent | 15×7 | 6 | 0,135 | 0,127 | 0,127 |
| | acl_dne | 37×7 | 7 | 0,135 | 0,119 | 0,118 |
| | | 150×7 | | | | |
| | Index | | | 0,183 | 0,141 | 0,142 |

Tabela 4.4 – Comparação dos métodos de simulação das corujas manchadas

| | | | | Probabilidades de decisão | | |
|---|---|---|---|---|---|---|
| Coruja | Variáveis | n | Escolhas | Newton-Raphson | Gauss-Laguerre | Hardy |
| 18 | acl_ed | 40×8 | 1 | 0,126 | 0,123 | 0,123 |
| | hgt_cent | 36×8 | 2 | 0,127 | 0,123 | 0,123 |
| | age_cent | 71×8 | 3 | 0,124 | 0,123 | 0,123 |
| | acl_p2 | 68×8 | 4 | 0,123 | 0,123 | 0,123 |
| | tba_cent | 26×8 | 5 | 0,128 | 0,123 | 0,123 |
| | acl_dne | 21×8 | 6 | 0,126 | 0,124 | 0,124 |
| | slp_cent | 74×8 | 7 | 0,126 | 0,123 | 0,123 |
| | drd_cent | 103×8 | 8 | 0,123 | 0,124 | 0,124 |
| | | 439×8 | | | | |
| | Index | | | 0,143 | 0,176 | 0,178 |
| 19 | wwb_cent | 40×7 | 1 | 0,139 | 0,128 | 0,127 |
| | hgt_cent | 65×7 | 2 | 0,077 | 0,132 | 0,132 |
| | acl_p2 | 30×7 | 3 | 0,139 | 0,130 | 0,129 |
| | acl_ed | 89×7 | 4 | 0,131 | 0,127 | 0,126 |
| | slp_cent | 40×7 | 5 | 0,156 | 0,128 | 0,126 |
| | drd_cent | 34×7 | 6 | 0,126 | 0,127 | 0,126 |
| | acl_mps | 52×7 | 7 | 0,154 | 0,128 | 0,127 |
| | | 350×7 | | | | |
| | | Index | | 0,172 | 0,120 | 0,121 |
| 20 | acl_p4 | 6×7 | 1 | 0,154 | 0,129 | 0,127 |
| | acl_p2 | 18×7 | 2 | 0,174 | 0,119 | 0,117 |
| | age_cent | 8×7 | 3 | 0,161 | 0,109 | 0,107 |
| | drd_cent | 2×7 | 4 | 0,110 | 0,146 | 0,145 |
| | hgt_cent | 15×7 | 5 | 0,165 | 0,125 | 0,123 |
| | slp_cent | 8×7 | 6 | 0,167 | 0,109 | 0,107 |
| | acl_ed | 17×7 | 7 | 0,169 | 0,126 | 0,124 |
| | | 74×7 | | | | |
| | Index | | | 0,163 | 0,169 | 0,170 |
| 21 | acl_ed | 29×9 | 1 | 0,106 | 0,113 | 0,114 |
| | acl_p4 | 37×9 | 2 | 0,104 | 0,103 | 0,104 |
| | acl_dne | 27×9 | 3 | 0,110 | 0,119 | 0,120 |
| | wwb_cent | 20×9 | 4 | 0,109 | 0,113 | 0,114 |
| | drd_cent | 58×9 | 5 | 0,108 | 0,106 | 0,106 |
| | acl_p2 | 39×9 | 6 | 0,109 | 0,116 | 0,117 |
| | slp_cent | 31×9 | 7 | 0,108 | 0,112 | 0,112 |
| | rwd_cent | 35×9 | 8 | 0,109 | 0,110 | 0,110 |
| | pww_cent | 38×9 | 9 | 0,109 | 0,109 | 0,110 |
| | | 314×9 | | | | |
| | Index | | | 0,121 | 0,116 | 0,117 |

Tabela 4.5 – Comparação dos métodos de simulação das corujas manchadas

| | | | | Probabilidades de decisão | | |
|---|---|---|---|---|---|---|
| Coruja | Variáveis | n | Escolhas | Newton-Raphson | Gauss-Laguerre | Hardy |
| 25 | slp_cent | 5×8 | 1 | 0,120 | 0,135 | 0,133 |
| | acl_p2 | 24×8 | 2 | 0,130 | 0,136 | 0,130 |
| | acl_p4 | 13×8 | 3 | 0,110 | 0,119 | 0,120 |
| | wwb_cent | 13×8 | 4 | 0,129 | 0,144 | 0,138 |
| | prw_cent | 22×8 | 5 | 0,134 | 0,146 | 0,139 |
| | drd_cent | 49×8 | 6 | 0,083 | 0,090 | 0,082 |
| | tba_cent | 4×8 | 7 | 0,115 | 0,145 | 0,136 |
| | age_cent | 24×8 | 8 | 0,097 | 0,123 | 0,118 |
| | | 154×8 | | | | |
| | Index | | | 0,167 | 0,200 | 0,204 |
| 26 | acl_p3 | 9×8 | 1 | 0,121 | 0,096 | 0,093 |
| | acl_ed | 18×8 | 2 | 0,122 | 0,091 | 0,089 |
| | slp_cent | 6×8 | 3 | 0,084 | 0,060 | 0,058 |
| | acl_dne | 29×8 | 4 | 0,089 | 0,083 | 0,079 |
| | wwb_cent | 23×8 | 5 | 0,122 | 0,090 | 0,087 |
| | phw_cent | 6×8 | 6 | 0,083 | 0,062 | 0,061 |
| | acln2 | 2×8 | 7 | 0,093 | 0,084 | 0,081 |
| | age_cent | 7×8 | 8 | 0,108 | 0,088 | 0,085 |
| | | 100×8 | | | | |
| | Index | | | 0,070 | 0,100 | 0,104 |

Tabela 4.6 – Variáveis para as corujas manchadas selecionadas

| Variáveis | | Escolhas | Freqüências |
|---|---|---|---|
| wwb_cent | (Área basal Whitewood (pés/acre)) | 1 | 757 |
| acl_p2 | (Proporção da zona tampão de bosque de idade entre 6 e 20 anos) | 2 | 416 |
| drd_cent | (Distância à estrada principal (em pés)) | 3 | 310 |
| acl_dne | (Distância da borda mais próxima (em pés)$^{(2)}$) | 4 | 410 |
| acl_p3 | (Proporção da zona tampão de bosque de idade entre 21 e 40 anos) | 5 | 733 |
| acl_mps | (Tamanho médio da mancha da zona tampão (em pés)) | 6 | 296 |
| slp_cent | (Porcentagem de inclinação do terreno) | 7 | 238 |
| hgt_cent | (Altura média das árvores no bosque (em pés)) | 8 | 402 |
| acl_pscv | (Porcentagem do coeficiente de variação do tamanho da mancha) | 9 | 390 |
| rwd_cent | (Área basal Redwood ( pés/acre)) | 10 | 1164 |
| acl_p4 | (Proporção da zona tampão de bosque de idade entre 41 anos ou mais) | 11 | 1206 |
| Número de observações | | | 6322 |

A figura (4.2) apresenta os ajustes do modelo usando a distribuição binomial e função de ligação logit com as variáveis selecionadas.

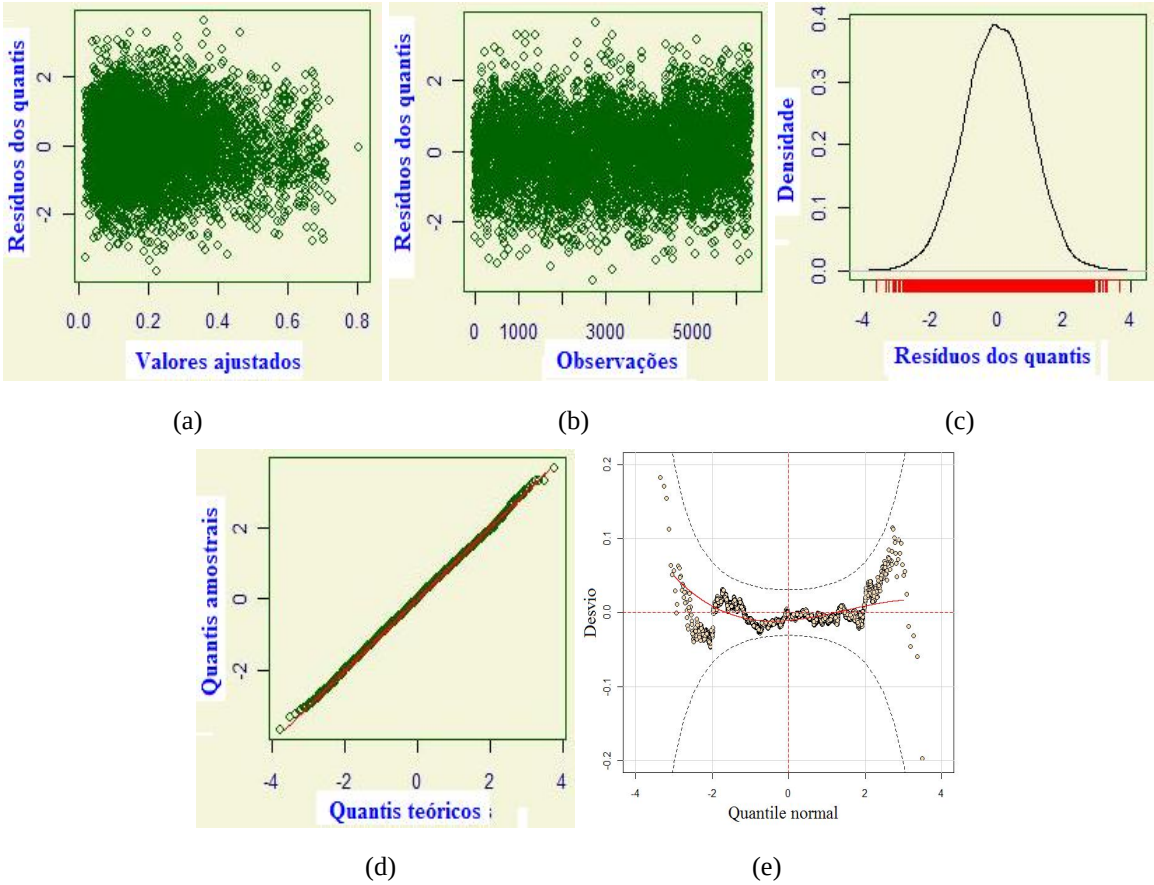



Figura 4.2 – Análises das corujas manchadas selecionadas, usando a distribuição Binomial

Tabela 4.7 – Variáveis selecionadas para as corujas manchadas, método Newton-Raphson

| Parâmetros | gl | Estimativas | Desvios Padrão | Aprox Pr > \|t\| |
|---|---|---|---|---|
| tacln_L1 | 1 | 0,0000136 | 8,3395 e-6 | 0,1041 |
| INC_L2G1C1 | 1 | 314,4346 | 195,9997 | 0,1087 |
| INC_L2G1C2 | 1 | 250,9493 | 155,3307 | 0,1062 |
| INC_L2G1C3 | 1 | 250,1283 | 155,3298 | 0,1073 |
| Likelihood Ratio (R) | 3161,5 | | | $2 * (\log L - \log L0)$ |
| Upper Bound of R (U) | 30319,0 | | | $-2 * \log L0$ |
| Cragg-Uhler 1 | 0,3935 | | | $1 - \exp(-R/N)$ |
| Cragg-Uhler 2 | 0,3968 | | | $\frac{(1-\exp(-R/N))}{(1-\exp(-U/N))}$ |
| Estrella | 0,4103 | | | 1 - $(1 - R/U)^{(U/N)}$ |
| Adjusted Estrella | 0,4101 | | | 1 - $(1 - R/U)^{(U/N)}$ |
| McFadden's LRI | 0,1043 | | | R/U |

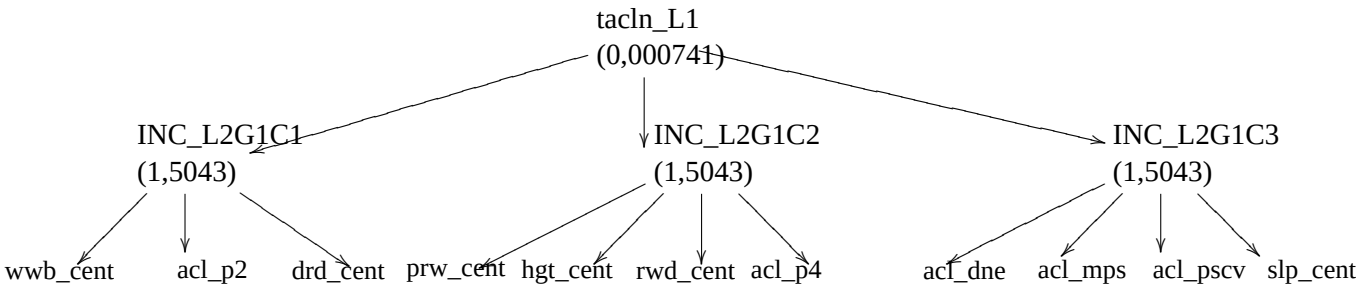



Figura 4.3 – Estrutura do Modelo logit encaixado para as corujas manchadas, usando as estimativas do método Newton-Raphson

Observa-se as análises das variáveis das corujas manchadas, selecionadas nas tabelas(4.2, 4.3, 4.4, 4.5) em que:

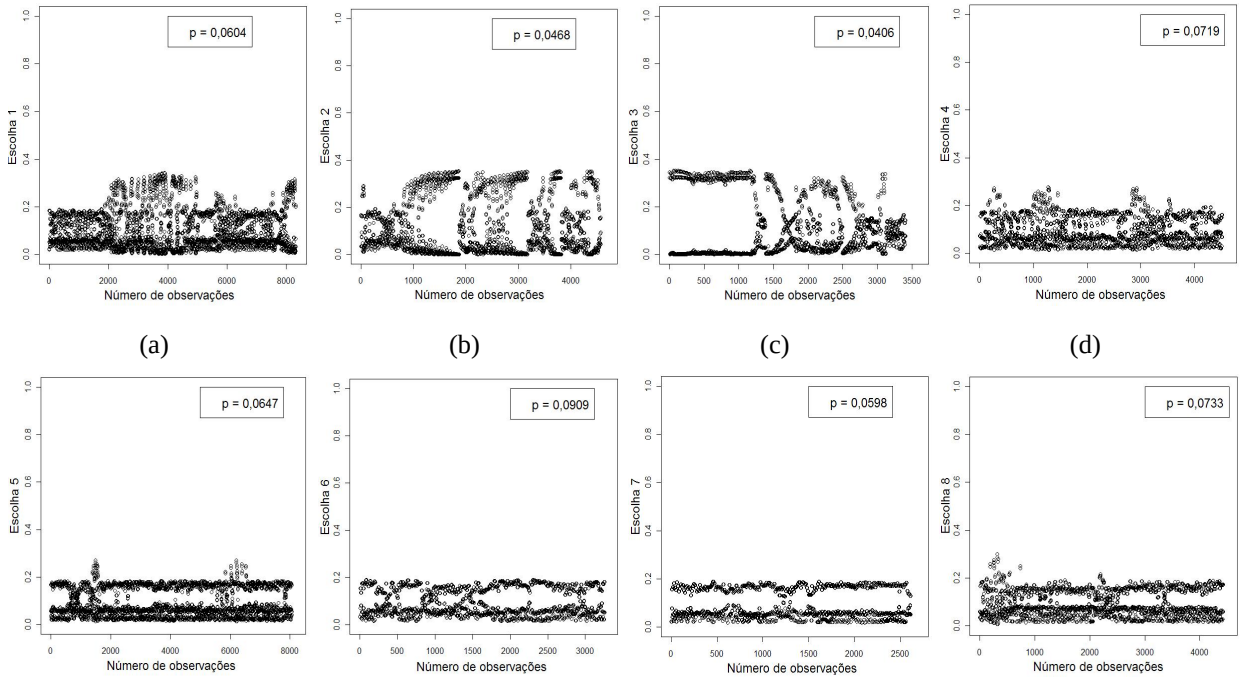
- Apresenta-se as análises das Corujas manchadas, com as 11 variáveis, selecionadas em que usou-se a distribuição binomial e a função de ligação logit, como observa-se nos gráficos do anexo, a figura (4.2) apresentando em (a), (b), uma dispersão dos resíduos dos quantis, observando-se que o modelo aceita dependência e é robusto, no (c) observa-se que os dados apresentam uma moda, no gráfico (d) observa-se o ajuste do modelo, no gráfico (e) ,“worm plot”, podemos observar uma cadeia contínua de pontos consecutivos (BUUREN; FREDRIKS, 2001).

- Observa-se o modelo logit encaixado type = nlogit, com a matriz de covariância usando a matriz inversa hessiana $COVEST = HESSIAN$, quando o método de Newton Raphson é usado. As estimativas são apresentadas na tabela (4.7), e na figura (4.3), (nesta figura as variáveis são agrupadas usando todas as combinações possíveis das escolhas até obter melhores resultados), as estimativas dos parâmetros do encaixe e dos encaixes o desvio padrão e a probabilidade aproximada. Também observa-se as estatísticas $Cragg - Ulher1$ e $Cragg - Ulher2$, que são bem similares, e as estatísticas Estrella e Estrella ajustada, que são de semelhante interpretação as estatísticas $R^2$. Observa-se na figura (4.4) e a tabela (4.8), que a melhor probabilidade de decisão do método Newton-Raphson foi obtida pela escolha 6, que corresponde á variável $acl\_mps$ (Tamanho médio da mancha da zona tampão (em pés)), mas as probabilidades de decisão das escolhas são todas muito similares.

Tabela 4.8 – Comparação dos métodos de simulação das corujas manchadas

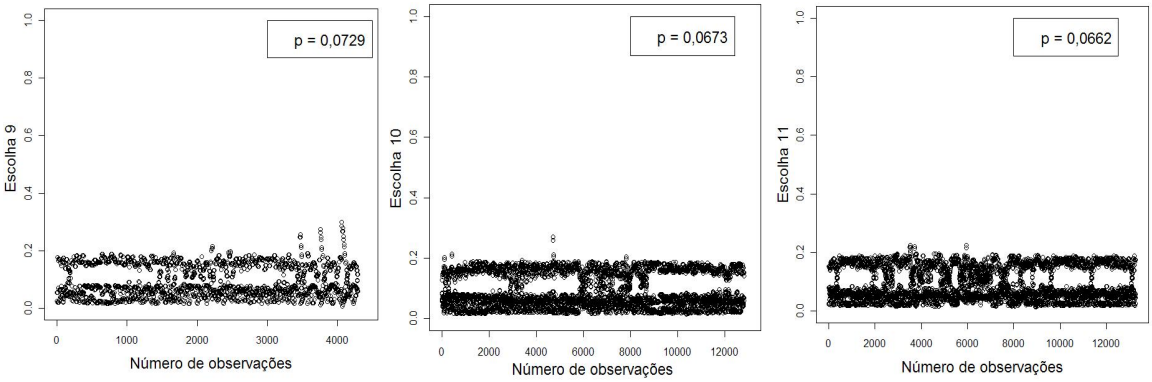
| | | | Probabilidades de decisão | | |
|---|---|---|---|---|---|
| Variáveis | n | Escolhas | Newton-Raphson | Gauss-Laguerre | Hardy |
| wwb_cent | 757×11 | 1 | 0,0604 | 0,0690 | 0,1330 |
| acl_p2 | 416×11 | 2 | 0,0468 | 0,0663 | 0,1300 |
| drd_cent | 310×11 | 3 | 0,0406 | 0,0659 | 0,1200 |
| acl_dne | 410×11 | 4 | 0,0719 | 0,0707 | 0,1380 |
| prw_cent | 733×11 | 5 | 0,0647 | 0,0701 | 0,1390 |
| acl_mps | 296×11 | 6 | 0,0909 | 0,0701 | 0,0827 |
| slp_cent | 238×11 | 7 | 0,0598 | 0,0704 | 0,1360 |
| hgt_cent | 402×11 | 8 | 0,0733 | 0,0770 | 0,1450 |
| acl_pscv | 390×11 | 9 | 0,0729 | 0,0743 | 0,1520 |
| rwd_cent | 1164×11 | 10 | 0,0673 | 0,0704 | 0,1180 |
| acl_p4 | 1206×11 | 11 | 0,0662 | 0,0699 | 0,1630 |
| | 6322×8 | | | | |
| Index | | | 0,2295 | 0,2316 | 0,2549 |

No index da razão da verossimilhança ajustado, da tabela (4.8), L(M) é o valor do modelo do logaritmo da verossimilhança, K o número de parâmetros, e L(C) o valor do logaritmo da verossimilhança do modelo nulo.




(a) (b) (c) (d)

(e) (f) (g) (h)

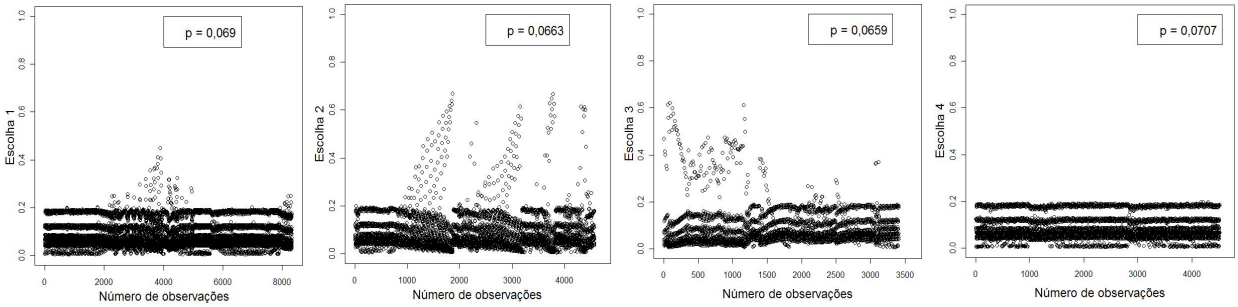



(i) (j) (k)

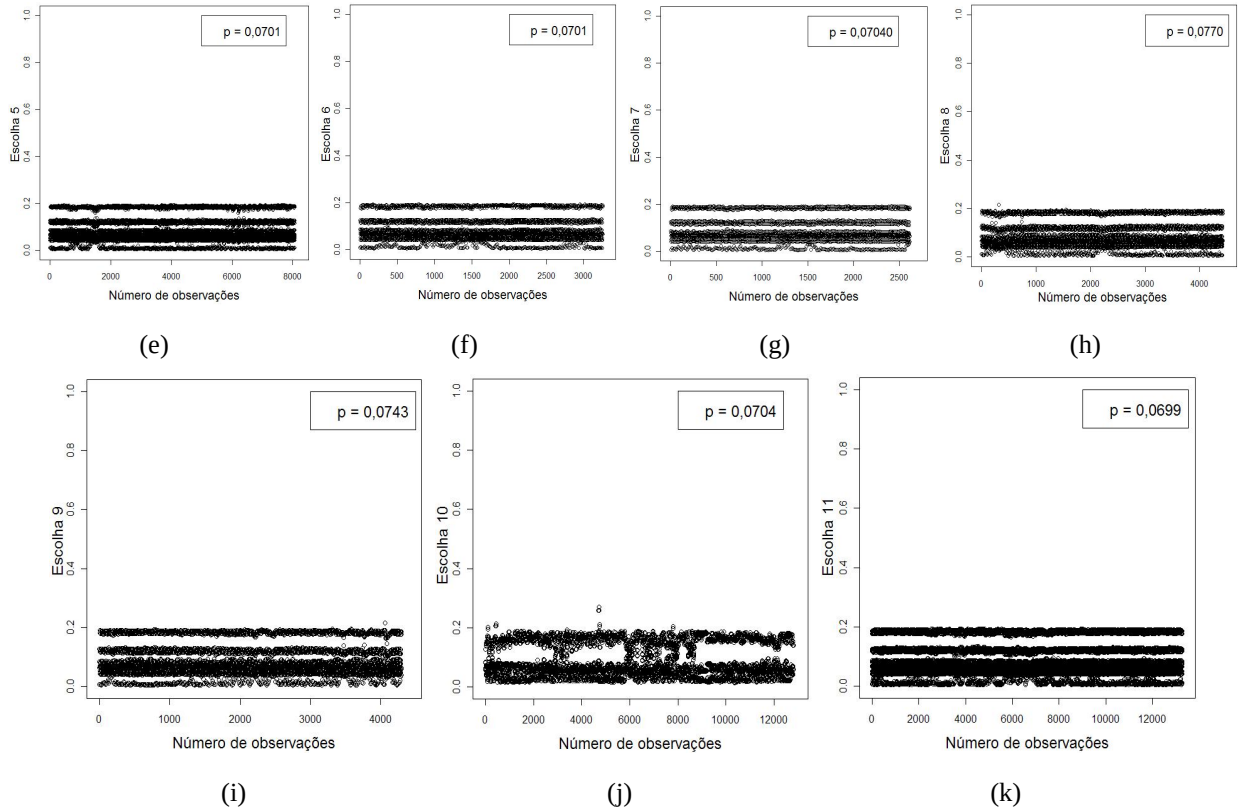
Figura 4.4 – Probabilidades de decisão das corujas manchadas selecionadas, usando o método Newton-Raphson

Tabela 4.9 – Variáveis selecionadas para a coruja manchada 1, Métodos Gauss-Laguerre e Hardy

| | | | Métodos Gauss-Laguerre | | Método Hardy | | |
|---|---|---|---|---|---|---|---|
| Parâmetros | gl | Estimativas | Desvios Padrão | Aprox Pr > \|t\| | Estimativas | Desvios Padrão | Aprox Pr > \|t\| |
| tacln | 1 | 0,000429 | 0,0000130 | <,0001 | 0,000433 | 0,0000159 | <,0001 |
| SCALE2 | 1 | 1,2585 | 0,0343 | <,0001 | 1,2778 | 0,0384 | <,0001 |
| SCALE3 | 1 | 3,6093 | 0,2309 | <,0001 | 3,6943 | 0,2504 | <,0001 |
| SCALE4 | 1 | 1,4874 | 0,0453 | <,0001 | 1,5081 | 0,0479 | <,0001 |
| SCALE5 | 1 | 1,0530 | 0,0301 | <,0001 | 1,0469 | 0,0309 | <,0001 |
| SCALE6 | 1 | 1,5744 | 0,0488 | <,0001 | 1,6125 | 0,0525 | <,0001 |
| SCALE7 | 1 | 1,7293 | 0,0557 | <,0001 | 1,7816 | 0,0604 | <,0001 |
| SCALE8 | 1 | 1,4065 | 0,0420 | <,0001 | 1,4265 | 0,0445 | <,0001 |
| SCALE9 | 1 | 1,4180 | 0,0424 | <,0001 | 1,4392 | 0,0450 | <,0001 |
| SCALE10 | 1 | 0,8008 | 0,0228 | <,0001 | 0,7895 | 0,0230 | <,0001 |
| SCALE11 | 1 | 0,7751 | 0,0221 | <,0001 | 0,7642 | 0,0223 | <,0001 |
| Index | | 0,2316 | | | 0,2549 | $1-\left(\frac{(\log L-K)}{\log LI}\right)^{\left(\frac{-2}{N*\log LI}\right)}$ | |




(a) (b) (c) (d)




(e) (f) (g) (h)

(i) (j) (k)

Figura 4.5 – Probabilidades de decisão das corujas manchadas selecionadas, usando o método Gauss-Laguerre

- Na tabela (4.9) usando o método de integração da quadratura gaussiana, Gauss-Laguerre e o de integração adaptativa de Hardy, podemos observar as estimativas e os desvios padrão dos métodos e observamos significância de todas as variáveis. Também observamos que na tabela (4.8) a melhor probabilidade de decisão do método Laguerre corresponde a escolha 8 e ao método de Hardy a escolha 11, que correspondem as variáveis $hgt\_cent$ (Altura média das árvores no bosque(em pés)), $acl\_p4$ (Proporção da zona tampão de bosque de idade entre 41 anos ou mais) (figuras (4.5)), mas as probabilidades de decisão são todas similares.

- Como observamos na tabela (4.8) o melhor Index de razão de verossimilhança corresponde ao Método de integração adaptativa Hardy.

Conclui-se que os resultados dos três métodos de simulação, método Newton-Raphson, método de quadratura gaussiana, Gauss-Laguerre e o método de integração adaptativa de Hardy, em relação as preferências de habitat das 11 corujas selecionadas, e usando a estrutura do modelo logit encaixado, como observa-se na figura (4.3), as variáveis do habitat selecionadas, usando como critério de seleção o Index da razão de verossimilhança, assim, as variáveis mas preferidas para o habitat das corujas manchadas selecionadas são a continuação,

- $acl\_p4$ (Proporção da zona tampão de bosque de idade entre 41 anos ou mais).

- $hgt\_cent$ (Altura média das árvores no bosque(em pés)).

- $acl\_mps$ (Tamanho médio da mancha da zona tampão (em pés)).

Considerando-se estas variáveis importantes para o habitat da coruja manchada (*Strix Occidentalis*).

**Referências**

ALDRICH, J. H.; NELSON, F. D. **Linear probability, logit, and probit models**. 2nd ed. Beverly Hills:Sage University papers, 1984. 95 p.

BHAT, C. A heteroscedastic extreme value model of intercity travel mode choice. **Transportation Research B**, Oxford, v. 29, n. 6, p. 471-483, 1995.

BUUREN V. S.; FREDRICKS M. Worm plot: a simple diagnostic device for modelling growth reference curves. **Statistic in Medicine**, Malden, v.20, p. 1259-1277, 2001.

BUTLER, J.S.; MOFFITT, R. A computationally efficient quadrature procedure for the one factor multinomial probit model. **Econometrica**, Menasha, v. 50, n. 3, p. 761-764, 1982.

CHINTAGUNTA, P.K.; JAIN, D.C.; VILCASSIM, N.J. Investigating heterogeneity in brand preferences in logit models for panel data. **Journal Marketing Research**, Chicago, v. 28, p. 417-428, 1991.

DAGANZO, C. **Multinomial probit:** economic theory, econometrics, and mathematical economics. New York, Academic Press. 1979. 222 p.

DALY, A.; ZACHARY, S. Improved multiple choice models. In: Henscher, A.J.; Dalvi, M.Q. (Ed.) **Determinants for Travel choice**, Westmead: Saxon House, 1978. 240 p.

ESTRELLA, A. A new measure of fit for equations with dichotomous dependent variables. **Journal of Business and Economic Statistic**, Washington, v. 16, n. 2, p. 198-205, 1998.

DHRYMES, J. Limited dependent variables. **Handbook of Econometrics**, Amsterdam, v.3, p. 1567-1631, 1996.

FORSMAN, E. D.; MESLOW, E.; WIGHT, H. M. Distribution and biology of the spotted owl in Oregon. **Wildlife Monographs**, Bethesda, v. 87, p. 1-64, 1984.

HERRIGES, J.; KLING, C. Testing the consistency of nested logit models with utility maximization. **Economics Letters**, Amsterdam, v. 50, p. 33-39, 1996.

JOHNSON, N.; KOTZ, S. **Distributions in statistical**: continuous univariate distribution. New York: Wiley Sons, 1970. chap. 21, p. 273-295.

KLING, C.; HERRIGES, J. An empirical investigation of the consistency of nested logit models with utility maximization. **American Journal of Agricultural Economics**, Ames, v. 77, p. 875-884, 1995.

LAYMON, S.A.; SALWASSER, H.; BARRETT, R.H. **Habitat suitability index models:** spotted owl. Washington, D.C: United States Fish and Wildlife Service, Biological Services Program. 1985. (Biological Report 82(10.113)).

MARSCHAK, J. Binary choice constraints on random utility indications. In: Arrow, K. **Stanford Symposium on mathematical methods in the social science**. Stanford:Stanford University Press, 1960. p. 312-329.

McCRACKEN, M.L.; MANLY, B.F.J.; VANDER HEYDEN, M. The use of discrete-choice models for evaluating resource selection. **Journal of Agricultural, Biological and Environmental Statistics**, Alexandria, v. 3, n. 3, p. 268-279, 1998.

McDONALD, L.L.; MANLY, B.F.J.; RALEY, C. M. Analyzing foraging and habitat use through selection functions. **Studies in Avion Biology**, Los Angeles, v. 13, p. 325-331, 1990.

McFADDEN, D. Conditional logit analysis of qualitative choice behavior. In: **Frontiers in econometrics**. (Ed), Academic Press: New York, p. 105-142. 1974.

McFADDEN, D. Modeling the choice of residential location. Amsterdam: **spatial interaction theory and residential location**, 1978. p.75-96.

McKELVEY, RD; ZAVOINA, W. A statistical model for the analysis of ordinal level dependent variables. **Journal of Mathematical Sociology**, New York, v. 4, p. 103-120. 1975.

PRESS, W.H; FLANNERY, B. P.; TEUKOLSKY, S. A; VETTERLING, W. T. **Numerical recipes**. Cambridge: Cambridge University Press, 1986. 818 p.

**R**: a language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0, Vienna, Austria, 2006, disponível em: <http://www.R-project.org>. Acesso em: 01 maio. 2006.

SAS Institute Inc.**SAS/STAT User's Guide, Version 9.1**. Carry, 2002-2003.

STROUD A. H; SECREST D.**Gaussian quadrature formulas**. Englewood: Prentice Hall. 1966. 374 p.

THURSTONE, L. A law of comparative judgement. **Psychological Review**, Princeton, v. 34, 1927. p. 273-286.

TRAIN, K. **Discrete choice methods with simulation**. Cambridge:Cambridge University Press, 2003. 334 p.

TVERSKY, A. Elimination by aspects: a theory of choice. **Psychological Review**, Princeton, v. 79, p. 281-299, 1972.

WILLIAMS, H. On the formation of travel demand models and economic evaluation measures of user benefits. **Environment and Planning A**, London, v. 9, p. 285-344, 1977.

# APÊNDICES

```
#################################################################
# Os PROGRAMAS PARA A ELABORAÇÃO DO CAPITULO 1,  foram feitos no
# código de programação Fortran e o código executável  estará
# disponível, na página do professor Carlos Tadeu dos Santos
# Dias ESALQ/USP.
#(http://www.lce.esalq.usp.br/tadeu.html
#(Teses e Dissertações de orientados))
#################################################################
# A1- Manual do programa de regressão logística (LOGREG)
#################################################################
# BASE DE DADOS DA AMOSTRA DE 390 OBSERVAÇÕES SELECIONADAS PARA
# AS ANÁLISES DE REGRESSÃO LOGÍSTICA
#####################################################################
USED owl trials aclc2 aclc3 aclc4 acl_ed    hgt_cent  slp_cent
 1   1    1  0    0  0    183.483   5          5
 0   1    1  1    0  0     77.862  44.382     25
                        ...
 0  26    1  1    0  0     55.519  19.593     46
```

The Computer Program LOGREG

The computer program LOGREG can be used to fit a variety of logistic regression models for proportion data, allowing for the effects of both quantitative variables (such as age in years) and factors (such as species of plant). If factors are considered then the user of LOGREG must set up appropriate 0-1 dummy variables. Fitted models always include a constant term.

According to the logistic assumption, the probability of a 'success' for cases in the ith of several groups is

$$p_i = \frac{\exp(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_p x_{ip})}{\boldsymbol{\beta}_0 + \exp(\boldsymbol{\beta}_1 x_{i1} + \ldots + \boldsymbol{\beta}_p x_{ip})}$$

This probability can then be estimated by knowing the number of 'successes' that occur with n trials for the ith group.

Input

Two types of input to the program are allowed:

(a) Input can be directly from the keyboard. In that case the user is first asked for the number of groups (proportions) and the number of predictor variables (X variables).
Next the numbers of "successes" and the number of trials (Y and n) are requested for the cases in turn. Finally, the values of predictor variables are requested, in the order X1, X2, ..., Xp. The analysis is then conducted, with output as shown in the example below.

(b) Input can be from a text file. In this case the first line of the file must contain a title and the second line must contain two numbers, separated by one or more spaces. These are the number of groups and the number of predictor variables. The rest of the file contains the number of 'successes' (Y), the number of trials (n), and the values of predictor variables (X1, X2, ..., Xp), one group at a time. The ASCII file can be set up in a text editor, a word

processor, or in a spreadsheet. The data should be set up as needed and then printed to an ASCII file.

## Output

In addition to appearing on the screen, output it is sent to the file LOGREG.OUT for later printing if desired.

## Example

This example is concerned with the site preference of two species of lizard, as shown in the following table of frequencies. The logistic regression model fits the proportion of species A present under various conditions. Only the main effects of the factors perch height, perch diameter, sun and time of day are allowed for.

| Perch height | Perch diameter | Sun or shade | Time of day | Species A | Total lizards |
|---|---|---|---|---|---|
| low | low | sun | early | 20 | 22 |
| high | low | sun | early | 13 | 13 |
| low | high | sun | early | 8 | 11 |
| high | high | sun | early | 6 | 6 |
| low | low | shade | early | 34 | 45 |
| high | low | shade | early | 31 | 36 |
| low | high | shade | early | 17 | 32 |
| high | high | shade | early | 12 | 13 |
| low | low | sun | mid-day | 8 | 9 |
| high | low | shade | mid-day | 55 | 59 |
| low | high | shade | mid-day | 60 | 92 |
| high | high | shade | mid-day | 21 | 26 |
| low | low | sun | late | 4 | 4 |
| high | low | sun | late | 12 | 12 |
| low | high | sun | late | 5 | 8 |
| high | high | sun | late | 1 | 2 |
| low | low | shade | late | 18 | 28 |
| high | low | shade | late | 13 | 16 |
| low | high | shade | late | 8 | 16 |
| high | high | shade | late | 4 | 8 |

Note that there were no lizards in the sun on large diameter high perches, at mid-day.
The input file for this example is as shown below, with 0-1 variables set up to account for the four factors, but no interactions.

```
                    HABITAT CHOICE OF LIZARDS
23 5
20 22 0 0 0 0 0
13 13 1 0 0 0 0
 8 11 0 1 0 0 0
 6  6 1 1 0 0 0
34 45 0 0 1 0 0
31 36 1 0 1 0 0
17 32 0 1 1 0 0
12 13 1 1 1 0 0
 8  9 0 0 0 1 0
 8  8 1 0 0 1 0
```

```
 4  5 0 1 0 1 0
69 89 0 0 1 1 0
55 59 1 0 1 1 0
60 92 0 1 1 1 0
21 26 1 1 1 1 0
 4  8 0 0 0 0 1
12 12 1 0 0 0 1
 5  8 0 1 0 0 1
 1  2 1 1 0 0 1
18 28 0 0 1 0 1
13 16 1 0 1 0 1
 8 16 0 1 1 0 1
 4  8 1 1 1 0 1
The output follows, with comments added in italics.
***** LOGISTIC REGRESSION *****
==============================================================
Problem title: HABITAT CHOICE OF LIZARDS
Number of X variables = 5 Number of groups = 23
Initial approximations for parameters
b0 1.176 b1 .000 b2 .000 b3 .000 b4 .000  b5 .000

The 5 0-1 variables represent the perch height (X1), the perch
diameter (X2), sun or shade (X3), and the three times of day
(X4 and X5).

Chi-squared values for null model with zero coefficients for X variables:

Pearson = 61.35, Log-likelihood = 70.10, with 22 df.

Iteration: 1 Initial Deviance = 70.102
Fraction of corrections made = 1.00000 New Deviance = 16.822

Iteration: 2 Initial Deviance = 16.822
Fraction of corrections made = 1.00000 New Deviance = 14.228

Iteration: 3 = Initial Deviance = 14.228
Fraction of corrections made = 1.00000 New Deviance = 14.205

Iteration: 4 Initial Deviance = 14.205
Fraction of corrections made = 1.00000 New Deviance = 14.205

Iteration: 5 Initial Deviance = 14.205
Fraction of corrections made = 1.00000 New Deviance = 14.205
Five iterations are required.

Final Standard
Parameter estimate error Ratio
==============================================
b0  .1945E+01 .3415E+00 5.69  Constant
b1  .1130E+01 .2571E+00 4.40  Perch height parameter
b2 -.7626E+00 .2113E+00 -3.61 Perch diameter parameter
b3 -.8473E+00 .3224E+00 -2.63 Shade (not sun) parameter
b4  .2271E+00 .2502E+00 .91   Mid-day parameter
b5 -.7368E+00 .2990E+00 -2.46 Late parameter

Group Y   N   E(Y)  Chi-sq   Group  Y   N   E(Y)  Chi-sq
=================================== ===================================
1     20  22  19.2   .24      2     13  13  12.4   .60
3      8  11   8.4   .09      4      6   6   5.5   .59
5     34  45  33.7   .01      6     31  36  32.5   .71
```

```
7     17  32  18.7   .35      8     12  13  10.6  1.05
9      8   9   8.1   .01     10      8   8   7.7   .29
11     4   5   4.0   .00     12     69  89  70.3   .12
13    55  59  54.3   .10     14     60  92  58.6   .09
15    21  26  22.0   .27     16      4   8   6.2  3.29
17    12  12  10.9  1.16     18      5   8   4.9   .01
19     1   2   1.7  1.52     20     18  28  16.5   .33
21    13  16  13.1   .00     22      8  16   6.4   .66
23     4   8   5.4  1.11
Goodness of fit: Pearson = 12.59, Log-likelihood = 14.20 , df = 17
The model is a good fit.

COVARIANCE MATRIX FOR PARAMETERS
b0  .117E+00
b1 -.157E-01  .661E-01
b2 -.216E-01  .447E-02  .446E-01
b3 -.825E-01 -.201E-02  .208E-03  .104E+00
b4 -.227E-01  .356E-02 -.432E-02 -.169E-01 .626E-01
b5 -.405E-01 -.303E-02  .220E-02  .361E-02 .363E-01 .894E-01

CORRELATION MATRIX FOR PARAMETERS
b0 1.000
b1 -.179 1.000
b2 -.299  .082  1.000
b3 -.750 -.024  .003  1.000
b4 -.266  .055 -.082  -.209 1.000
b5 -.397 -.039  .035   .037  .486 1.000
The model seems to fit, with no large residuals. All the parameters
 seem important except that the
proportion of species A at mid-day does not seem much different from
the proportion in the early part
of the day.

library(MASS)
library("gamlss")

DaCoruja<-read.table("E:/AnaFinalSAS/owl390.txt",header=TRUE)
attach(DaCoruja)
DaCoruja
modelo1<-glm(USED~aclc2+aclc3+aclc4+acl_ed+hgt_cent+slp_cent,
data=DaCoruja,family=binomial(logit))
summary(modelo1)
anova(modelo1)

modelo1g<-gamlss(USED~aclc2+aclc3+aclc4+acl_ed+hgt_cent+slp_cent,
data=DaCoruja,family=BI())
summary(modelo1g)
plot(modelo1g)
wp(modelo1g)

#################################################################
# A2- Manual do programa do modelo de escolha discreta com
# uma escolha. (DISCCHSE)
#################################################################
# BASE DE DADOS PARA USAR O PROGRAMA DE MODELO DE ESCOLHA DISCRETA
# COM UMA ESCOLHA
##################################################################

Test owl 390
390 6
```

```
1  1 0 0 0 183.5  5      5
0  1 1 0 0  77.9 44.4 25
                ...
0 26 1 0 0  55.5 19.6 46
```

The Computer Program DISCCHSE

This program fits the discrete choice model to data using the principle of maximum likelihood. The situation is that one or more animals makes S independent choices of which resource unit to use. At the ith choice there are ni units that can be selected, in what is called the choice set. Only one of these units is chosen. Each unit is described by its values for p variables X1, X_2,..., X_p, and the probability that the jth unit is selected at the ith choice is

$$p_{ij} = \frac{\exp(\boldsymbol{\beta}_1 x_{ij1} + \boldsymbol{\beta}_2 x_{ij2} + \ldots + \boldsymbol{\beta}_p x_{ijp})}{\sum_{k=1}^{n_i} \exp(\boldsymbol{\beta}_1 x_{ik1} + \boldsymbol{\beta}_2 x_{ik2} + \ldots + \boldsymbol{\beta}_p x_{ikp}).}$$

where x_iku is the value of X_u for the kth unit that is available at the ith choice.

Data Input

The data takes the following format:

Line 1: A title for the data set.

Line 2: NUNIT, NX, where NUNIT is the total number of units for which data are available in all the choice sets, and NX is the number of X variables measured on units (called p above)..

Line 3: IND(1), SET(1), X(1,1), X(1,2), ..., X(1,NX), where X(1) = 1 if this unit is selected, or otherwise is 0, SET(1) is the number of the choice set that this unit is in, and X(1,1) to X(1,NX) are the values of the X variables (in order) for this unit.
...
Line NUNIT+2: IND(NUNIT), SET(NUNIT), X(NUNIT,1), X(NUNIT,2), ..., X(NUNIT,NX), where the terms have the same meaning as on line 3, but for the last resource unit in the data set.
Note that the number of choice sets in total (S above) is SET(NUNIT) because the results must be given for choice set 1, then choice set 2, and so on up to the last choice set. In other words, SET(1) = 1, SET(2) = 2, etc. Also, the number of units in each choice set can vary.
An example set of test data for the program is as follows, where there are 250 resource units altogether in 50 choice sets with five units in each set, and where two X variables are measured on each resource unit: Test Data for Discrete Choice

```
250 2
0 1 4.7 0.7
0 1 0.7 0.2
0 1 3.8 0.1
1 1 7.4 0.1
     .
     .
```

```
     .
0 50 4.8 5.5
0 50 0.5 6.7
0 50 2.7 5.8
```

Computer Output

The output obtained from the program with the above data are shown below. Comments are provided in italics on the right.

```
     &' ************************************************************'/
     &' *                        DISCCHSE                          *'/
     &' *                                                          *'/
     &' * A program for fitting the discrete choice model to data. *'/
     &' * This program is based on the program MAXLIK written by   *'/
     &' * T.E. Reed for fitting a general multinomial model.       *'/
     &' *                                                          *'/
     &' * There can be up to 50000 units, with up to 50 estimated  *'/
     &' * parameters.                                              *'/
     &' *                     Written by                           *'/
     &' *                   Sandra Vergara Cardozo                 *'/
     &' *                    Bryan F.J. Manly                      *'/
     &' *    ESALQ/USP   and Western EcoSystems Tecnology, Inc.    *'/
     &' *                                                          *'/
     &' *             Version 1.1, Dated January, 2007             *'/
     &' *                                                          *'/
     &' ************************************************************'//

Date: 28/ 1/2007 Time: 13:41
Data Title: Test Data for Discrete Choice
Number of Units = 250
Unit Use  Set X-Values
 1    0   1  4.70 0.70
 2    0   1  0.70 0.20
 3    0   1  3.80 0.10
 4    1   1  7.40 0.10
 5    0   1  2.60 0.80
 6    0   2  0.50 4.80
 7    0   2  4.70 5.40
 8    0   2  2.10 3.40
 9    1   2  8.90 1.10
10    0   2  9.60 6.20
11    0   3  3.70 6.20
12    1   3  7.70 7.00
13    0   3  4.60 5.30

.
.
.
247   1 50  8.60 2.00
248   0 50  4.80 5.50
249   0 50  0.50 6.70
250   0 50  2.70 5.80

                    *** NULL MODEL FIT ***
The null model has no parameters and gives all choices the same
probability

Log-likelihood = -80.472
Deviance = AIC = 160.944 with 250 df
Models allowing for selection related to the X-variables
```

```
are now fitted. This process can be repeated as much as
necessary. First, variables to be used are chosen.
Variables chosen for use are:
1
Just fitting variable 1 first.

                  *** SELECTION MODEL FIT ***

Parameters of the fitted model are the coefficients of the 1
X-variables.

                     Summary of Iterations

                        Initial Final

Iter   log-lik     log-lik    Halvings
1      -80.472     -60.504     1
2      -60.504     -59.680 1
3      -59.680     -59.669 1
4      -59.669     -59.669 2
Final Deviance = 119.34 with 249 df
AIC = 121.34

Final Standard Last
Param estimate error change
1 0.4030 0.0753 0.0002

Correlation matrix for parameters
1 1.000
Now asked for all (two) variables to be included in the model.

                 *** SELECTION MODEL FIT ***

Parameters of the fitted model are the coefficients of the 2
X-variables.
                       Summary of Iterations
                           Initial Final

Iter   log-lik  log-lik  Halvings
1      -80.472 -37.943    1
2      -37.943 -29.951    1
3      -29.951 -27.445    1
4      -27.445 -27.028    1
5      -27.028 -27.010    1
6      -27.010 -27.010    1
7      -27.010 -27.010    2
Final Deviance = 54.02 with 248 df
AIC = 58.02

Final Standard Last
Param estimate  error   change
1     0.9377    0.1992 -0.0006
2    -0.8611    0.1756  0.0004
Correlation matrix for parameters
1 1.000
2 -0.806 1.000
The true parameter values for these artificial data are +1 and
-1, so the estimates are quite reasonable at 0.9377 and -0.8611.
```

```
##############################################################
# B- A continuação da saída do programa RSPFSIM, feito na
# linguagem de programação Fortran, os programas foram feitos
# para o bootstrapping paramétrico e não-paramétrico, os
# programas geram o conjunto de dados e depois fazem as
# análises, encontram-se disponíveis na página do professor
# Carlos Tadeu dos Santos Dias ESALQ/USP.
#(http://www.lce.esalq.usp.br/tadeu.html
#(Teses e Dissertações de orientados))
##############################################################

 ************************************************************
 *                       RSPFSIM                            *
 *                                                          *
 * A program to simulate the estimation of a resource       *
 * selection probability function (RSPF) using data on a    *
 * set of units observed to be used and a large sample of   *
 * the other units.  It is assumed that the RSPF has the    *
 * logistic form                                            *
 *                                                          *
 *                    exp(b0+b1X1+ ... + bpXp)              *
 *         P(i) = ---------------------------------         *
 *                 1 + exp(b0 + b1X1 + ... + bpXp)          *
 *                                                          *
 * where X1 to Xp are variables that define the probability *
 * of use.  Estimation is by pseudo maximum likelihood,     *
 * with the assessment of standard errors by parametric     *
 * bootstrapping.                                           *
 *                                                          *
 * There can be up to 50000 units in the set of used units  *
 * plus the large sample of other units.  There can be up   *
 * to 100 X variables.                                      *
 *                                                          *
 *                       Written by                         *
 *                     Bryan F.J. Manly                     *
 *                  Sandra Vergara Cardozo                  *
 *  Western EcoSystems Technology Inc. and ESALQ/USP        *
 *            Version 1.1, Dated February, 2007             *
 ************************************************************

 Date: 19/ 2/2007  Time: 20:50


 Data Generation Details
 Number of units in the population (NUNIT) =   1000000
 Expected number of used units (N1) =   1000
 Expected number of unused units sampled (N2) =   1000
 Number of X variables (NX) =   4
 Random number seed =  30000

 Regression coefficients for X variables
  1.00  2.00  3.00  4.00

 Estimation of a Resource Selection Probability Function
 Number of units with data =  1908  Total number of units =  1000000
 Number of X variables =   4

 Input Data
               X Variable Values
  Case Used       1       2       3       4
```

```
    1    0   -0.35   -1.80   -0.62   -1.18
    2    0    1.05    0.56    0.89   -0.83
    3    0    0.32   -1.17   -0.12   -1.31
    4    0   -0.71    0.00   -2.04    0.46
                         .
                         .
                         .
 1906    1    0.65    1.67    1.20    1.94
 1907    1    2.30    2.18    0.15    2.67
 1908    1   -0.26    1.13    1.35    2.53

Sorted Data (Used Units First)
    1    1    0.32    1.94    2.07    2.82
    2    1   -0.04    1.17    1.74    2.02
    3    1    1.31    1.40    2.21    1.89
    4    1   -0.11    0.04    3.56    2.64
    5    1    0.44    0.70    2.62    1.57
    6    1    0.68    0.62    0.54    3.18
    7    1   -0.23    1.34    2.06    1.30
    8    1    1.57    2.30    1.12    2.73
    9    1   -0.90    0.92    1.17    2.04
   10    1    2.46    1.63    1.98    1.42
    .
    .
    .
 1907    0   -1.34   -1.81   -0.93    0.59
 1908    0    1.77   -0.77   -0.88   -1.46

Null Model for the RSPF
Data units recorded as used =  908
Data units not recorded as used = 1000
Total number of unused units =   999092
Probability of recorded use for all units = 0.0009
Residual df = 1907
Pseudo log-likelihood =  -7267.48  Pseudo-deviance =  14534.96

Fitted Model with Standard Deviations Based on Usual Likelihood Theory
Residual df = 1903
Pseudo log-likelihood =  -1506.84  Pseudo-deviance =   3013.68
Change in deviance= 11521.276  df =    4

Param  Estimate        SE
    1   -27.912     0.798
    2     1.966     0.087
    3     2.219     0.114
    4     6.552     0.208
    5     6.395     0.204

Bootstrap Results
      Change in
 Data  Deviance   Parameter Estimates
    1******************************218132.891221780.375278525.500
    2**********-87376.055 60941.387-12130.455 14076.127-14425.365
                                        ...
  499********** 48438.449-19066.703 -3074.976-33183.320 -7000.982
  500********** 18912.121-35603.355******************** -4780.709
```

```
##############################################################
# USANDO O PROGRAMA R PARA ELABORAÇÃO DO CAPITULO 3
# TODAS AS VARIABLES DAS CORUJAS
##############################################################
library(rpart)
library("gamlss")
library(MASS)
library(bpca)
library(graphics)
library(car)
library(vegan)
#####################################################################
# C1- Análises geral e probabilidade de decisão de todas as corujas
#####################################################################
# TODAS AS VARIABLES E TODAS AS CORUJAS
#####################################################################

fowl<- rpart(USED~
SPECENT2+SPECENT3+aclc2+aclc3
+aclc4+acln2+acln3+acln4+aclc2_acln2
+aclc2_acln3+aclc2_acln4+aclc3_acln2+aclc3_acln3+aclc3_acln4
+aclc4_acln2+aclc4_acln3+aclc4_acln4+age_cent+acl_p2
+acl_p3+acl_p4+acl_dne+acl_ed+acl_mps
+acl_np+acl_pd+acl_pscv+acl_te+drd_cent
+hgt_cent+phw_cent+prs_cent+prw_cent+pww_cent
+rwd_cent+wwb_cent+hwd_cent+res_cent+tba_cent
+slp_cent, data=owl)
summary(fowl)

par(mfrow=c(2,2), xpd=NA)
plot(fowl)
text(fowl, digits = 2,use.n=TRUE)
plotcp(fowl)
wp(fowl,ylim.all=0.4) # O figura worm plot
##############################################################
# Ajustando, os dados, e usando a distribuição
# Binomial e o link logit
##############################################################

fowli <- gamlss(USED~
SPECENT2+SPECENT3+aclc2+aclc3
+aclc4+acln2+acln3+acln4+aclc2_acln2
+aclc2_acln3+aclc2_acln4+aclc3_acln2+aclc3_acln3+aclc3_acln4
+aclc4_acln2+aclc4_acln3+aclc4_acln4+age_cent+acl_p2
+acl_p3+acl_p4+acl_dne+acl_ed+acl_mps
+acl_np+acl_pd+acl_pscv+acl_te+drd_cent
+hgt_cent+phw_cent+prs_cent+prw_cent+pww_cent
+rwd_cent+wwb_cent+hwd_cent+res_cent+tba_cent
+slp_cent,family=BI(), data=owl)
summary(fowli)

plot(fowli)
wp(fowli,ylim.all=0.4)
##############################################################
## C2 - Coruja 1
##############################################################

owl1<-(read.csv(file="F:/AnaFinalSAS/owl1/owl1all.csv",header=TRUE))
attach(owl1)
names(owl1)
```

```
library(rpart)
set.seed(123)
#inicialmente con todas as variáveis
fowl1 <- rpart(USED~
SPECENT2+SPECENT3+aclc2+aclc3
+aclc4+acln2+acln3+acln4+aclc2_acln2
+aclc2_acln3+aclc2_acln4+aclc3_acln2+aclc3_acln3+aclc3_acln4
+aclc4_acln2+aclc4_acln3+aclc4_acln4+age_cent+acl_p2
+acl_p3+acl_p4+acl_dne+acl_ed+acl_mps
+acl_np+acl_pd+acl_pscv+acl_te+drd_cent
+hgt_cent+phw_cent+prs_cent+prw_cent+pww_cent
+rwd_cent+wwb_cent+hwd_cent+res_cent+tba_cent
+slp_cent, data=owl1)
par(mfrow=c(2,2), xpd=NA)
plot(fowl1)
text(fowl1, digits = 2,use.n=TRUE)
plotcp(fowl1)

fowl1i <- gamlss(USED~
SPECENT2+SPECENT3+aclc2+aclc3
+aclc4+acln2+acln3+acln4+aclc2_acln2
+aclc2_acln3+aclc2_acln4+aclc3_acln2+aclc3_acln3+aclc3_acln4
+aclc4_acln2+aclc4_acln3+aclc4_acln4+age_cent+acl_p2
+acl_p3+acl_p4+acl_dne+acl_ed+acl_mps
+acl_np+acl_pd+acl_pscv+acl_te+drd_cent
+hgt_cent+phw_cent+prs_cent+prw_cent+pww_cent
+rwd_cent+wwb_cent+hwd_cent+res_cent+tba_cent
+slp_cent,family=BI(), data=owl1)
summary(fowl1i)

plot(fowl1i)
wp(fowl1i,ylim.all=0.4)
################################################################
# Com as variáveis selecionadas da Coruja manchada 1
################################################################
fowl1ib<-gamlss(USED~slp_cent+ acl_p2+acl_p3+ drd_cent+acl_p4 +hwd_cent
+acl_pscv+age_cent,family=BI(),data=owl1)
summary(fowl1ib)
plot(fowl1ib)
wp(fowl1ib)
list(fowl1ib)
################################################################
# C3-1 Programa da linguagem SAS
################################################################

Para a Coruja 1, com 8 escolhas,

data newdata1(keep=pid decision mode tacln);
      set owl1;
      array tvec{8} acln1 - acln8;
      retain pid 0;
      pid + 1;
      do i = 1 to 8;
         mode = i;
         tacln = tvec{i};
         decision = ( choice = i );
         output;
      end;
   run;
```

```
   proc print data=newdata1; # os dados pronosticados
   run;
################################################################
# Usando o método de otimização quasinewton e a matriz hessiana.
################################################################

proc mdc data=newdata1;
      model decision = tacln / type=clogit nchoice=8
         optmethod=qn covest=hess;
      id pid;
  output out=probdata pred=p;
   run;
#################################################################
# C3-1 Usando o método Newton Raphson
#################################################################

 proc mdc data=newdata1 outest=e1s;
      model decision = tacln /
            type=nlogit choice=(mode 1 2 3 4 5 6 7 8)
            spscale covest=hess;
      id pid;
      utility u(1,) = tacln;
      nest level(1) = (1 2 3 @ 1, 4 5 6  @ 2, 7 8 @ 3),
           level(2) = (1 2 3 @ 1);
   output out= owl1NR PRED= pdec1NR;
   run;
#####################################################################
# C3-2 O método de integração da quadratura gaussiana, Gauss-Laguerre
# com o valor extremo heterocedástico (hev).
#####################################################################

proc mdc data=newdata1;
      model decision = tacln / type=hev nchoice=8
            hev=(integrate=laguerre) covest=hess;
      id pid;
  output out= owl1L PRED= pdec1L;
run;
####################################################################
# C3-3 método de integração adaptativa no método de Hardy, com o
# valor extremo heterocedástico (hev),
####################################################################

 proc mdc data=newdata1;
      model decision =tacln / type=hev nchoice=8
         hev=(integrate=hardy) covest=hess;
      id pid;
    output out= owl1H PRED= pdec1H;
   run;
```

E assim sucessivamente ate a Coruja 26