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

Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (*Euschistus heros Fabricius*) da soja (*Glycine max L.*)

Raphael Antonio Prado Dias

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

Piracicaba
2014

Raphael Antonio Prado Dias
Licenciado em Matemática

Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (*Euschistus heros Fabricius*) da soja (*Glycine max L*)

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **CARLOS TADEU DOS SANTOS DIAS**

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

Piracicaba
2014

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Dias, Raphael Antonio Prado
Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (*Euschistus heros Fabricius*) da soja (*Glycine max L*) / Raphael Antonio Prado Dias.- - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2014.
97 p: il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2014.

1. Distribuições multivariadas 2. Poisson 3. Normal 4. Simulação I. Título

CDD 632.754
D541s

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte -O autor"

Dedicatória

A Deus,

Ao meu pai Edson Mesquita Dias,

À minha mãe Deuseli Aparecida Prado Dias,

Às minhas irmãs Aline e Ariane,

À minha namorada Jeniffer Roberta Pires Chagas.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Carlos Tadeu dos Santos Dias, pelo conhecimento compartilhado, ao apoio, a amizade, estímulo constante e pelas críticas e sugestões fundamentais para a realização deste trabalho.

Aos professores do programa de Pós-graduação em Estatística e Experimentação Agronômica, Dr. Paulo Justiniano Ribeiro Jr., Dra. Roseli Aparecida Leandro, Dra Clarice Garcia Borges Demétrio, Dr. César Gonçalves de Lima, Dr. Décio Barbin, Dr. Gabriel Adrián Sarriés, Dr. Sílvio Zocchi, Dra. Sônia Maria de Stefano Piedade, Dra. Taciana Villela Savian, Dr Idemauro Antonio Rodrigues de Lara, Dr. Cristian Marcelo Villegas Lobos por todos os ensinamentos e amizade.

Aos funcionários do Departamento de Ciências Exatas Solange Sabadin, Luciane Brajão, Mayara Segatto, Eduardo Bonilha, Jorge Wiendl e Rosni Honofre.

Aos amigos dos cursos de mestrado e doutorado do Departamento de Ciência Exatas da ESALQ/USP, em Especial a Adriana Maria Marques da Silva, Ana Patrícia Bastos Peixoto, Braga Júnior, Cássio Dessotti, Cristiane Rodrigues, Elizabeth Mie Hashimoto, Edilan Quaresma, Eduardo Monteiro, Epaminondas Couto, Everton Batista, Ezequiel Bautista, Fabiane Silva, Fernada Rizzato, Guilherme Biz, Gláucia Tatiana Ferrari, Keliny Martins, Kuang Hongyu, Josiane Rodrigues, Juliana Betine, Juliana Fachini, Luiz Roberto, Marcelino Rosa de Pascoa, Mariana Urbano, Marina Maestre, Maurício Lordelo, Patrícia Paranaíba, Pedro Henrique, Priscila Neves Faria, Renato Nunes Pereira, Renata Alcarde, Rodrigo Pescim, Ricardo Olinda, Simone Werner, Simone Sartório, Tiago Almeida de Oliveira, Tiago Santana, pelo auxílio em momentos de duvidas, e principalmente pela sincera amizade.

À minha namorada Jeniffer, sua mãe Deuzelena, suas irmãs Germana e Joyce e seus cunhados Renatinho e Daniel pela amizade e apoio em todos os momento.

Aos amigos de trabalho do Instituto Federal de Educação, Ciência e Tecnologia Sul de Minas Gerais pela amizade e apoio.

Aos amigos Marcos Roberto do Prado, Renato Moreira e Lucas Willian Mendes, a amizade e a boa convivência durante o período de república.

A Jordana Rebolla Martha pela prontidão em fazer a correção ortográfica.

Aos amigos da Associação de Pós-Graduação em especial Maísa, Sérgio e

Biba.

A Universidade de São Paulo pela oportunidade de realização do curso de pós-graduação.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES pelo apoio financeiro em forma de bolsa de estudos.

A todas as pessoas que de alguma forma contribuíram para minha formação profissional e para a realização deste trabalho.

SUMÁRIO

RESUMO 9
ABSTRACT 11
LISTA DE FIGURAS 13
LISTA DE TABELAS 15
1 INTRODUÇÃO 17
2 OBJETIVO 19
3 DESENVOLVIMENTO 21
3.1 Revisão Bibliográfica 21
3.1.1 Distribuição Uniforme 21
3.1.2 Distribuição Normal Univariada 21
3.1.3 Distribuição Poisson 22
3.1.4 Dados Multivariados 23
3.1.5 Distribuição Normal Multivariada 24
3.1.6 Geração de variáveis aleatórias correlacionadas 27
3.1.6.1 Geração de variáveis aleatórias de uma distribuição normal multivariada . . 27
3.1.6.2 Geração de variáveis aleatórias uniformes correlacionadas 29
3.1.6.3 Geração de variáveis aleatórias Poisson correlacionadas 29
3.1.7 Teste de Aderência 40
3.1.8 Simulação 41
3.1.9 Regressão Não-Linear 41
3.1.10 Regressão Linear Múltipla 49
3.1.11 Método dos Mínimos Quadrados 50
3.1.12 Teste t e teste F 51
3.1.13 Coeficiente de Correlação Múltiplo 51
3.1.14 Métodos de seleção de variáveis 52
3.1.15 Análise de resíduos 54
3.2 Material e Métodos 58
4 RESULTADOS 61
4.1 Resultados das Simulações 61
4.1.1 Médias e correlações iguais 61
4.1.2 Médias iguais e correlações diferentes 66

4.2 Exemplo de Aplicação 70
4.2.1 Análise da taxa de falsos alertas do aumento populacional de percevejos na *Glycine max L* (soja) 70
4.2.1.1 Regressão linear múltipla para a taxa de falsos alertas 75
4.2.1.2 Método de seleção de variáveis 75
4.2.1.3 Análise de Resíduos 75
4.2.1.4 Estimação dos parâmetros e análise de variância 78
5 CONCLUSÕES 79
REFERÊNCIAS 81
ANEXOS 85

RESUMO

Simulação estocástica de variáveis aleatórias Poisson correlacionadas: aplicação ao controle populacional do percevejo (*Euschistus heros Fabricius*) da soja (*Glycine max L.*)

A simulação de dados que seguem distribuição de Poisson é essencial em muitas aplicações reais de várias áreas, tais como saúde, marketing, ciências agronômicas, entre outras em que os dados são contagens multivariadas. Métodos de simulação atuais sofrem de limitações computacionais e restrições à estrutura de correlação e, portanto, são raramente usados. Neste trabalho propôs-se uma modificação do método NORTA para gerar dados com distribuição Poisson multivariada a partir de uma distribuição normal multivariada com matriz de correlações e vetor de médias pré estabelecidos. Como as distribuições Normal multivariada e univariada e a distribuição Poisson univariada já estão implementadas em softwares estatísticos, inclusive no R, implementou-se algumas linhas de código. Mostrou-se que o método funciona bem e é altamente preciso na geração de dados multivariados com distribuição marginais de Poisson, para diferentes estruturas de correlações (negativas e positivas e variando os valores) e para altos e baixos valores de médias. Mostrou-se as vantagens práticas da simulação de dados de Poisson multivariada sobre a normal multivariada na detecção da taxa de falsos alertas de super populações de percevejos, evidenciando que simulações inadequadas podem levar a excesso de falsos alertas. Uma vez que os dados seguem distribuição Poisson multivariada, a taxa de falsos alertas pode ser maior do que a imaginada. Essa taxa pode ser estimada por um modelo ajustado. A mesma técnica pode ser aplicada em diversos problemas de várias áreas do conhecimento.

Palavras-chave: Distribuições multivariadas; Poisson; Normal; Simulação

ABSTRACT

Stochastic simulation for correlated Poisson random variables: application to population control bedbug (*Euschistus heros Fabricius*) soy (*Glycine max L.*)

The simulation data that follow a Poisson distribution is essential in many real applications in various areas such as healthcare, marketing, agronomic sciences, among others that the data are multivariate counts. Current simulation methods suffer from limitations and constraints on computing correlation structure and are therefore seldom used. This paper proposed a modification of the NORTA method for generating data with multivariate Poisson distribution from a multivariate normal distribution with correlation matrix and vector of predetermined average. As the multivariate and univariate Normal distribution and univariate Poisson distribution are already implemented in statistical software, including R, was implemented just a few lines of code. It was shown that the method works well and is highly accurate in generating multivariate data with marginal Poisson distribution structures for different correlations (negative and positive values) and for high and low λ. Proved the practical benefits of the simulation data on the multivariate Poisson multivariate normal in the detection of super bugs populations, inadequate simulations can lead to excessive false alerts. Once the data are multivariate Poisson distribution, the rate of false alarms can be greater than the imagined. This rate can be estimated by an adjusted model. The same technique can be applied to many problems in various fields of knowledge.

Keywords: Multivariate distributions; Poisson; Normal; Simulation

LISTA DE FIGURAS

Figura 1 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0$ 26

Figura 2 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0,5$ . . 26

Figura 3 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0,9$ . . 27

Figura 4 - Distribuição normal bivariada com coeficiente de correlação $\rho = -0,5$. 27

Figura 5 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 1 61

Figura 6 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 2 62

Figura 7 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 5 63

Figura 8 - Frequências observadas e frequências esperadas dos valores simulados das variáveis aleatórias com distribuição Poisson e média 10 64

Figura 9 - Frequências observadas e frequências esperadas dos valores simulados das variáveis aleatórias com distribuição Poisson e média 20 64

Figura 10 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 40 65

Figura 11 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações 0,5; -0,5 e 0,25 66

Figura 12 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações 0,9; -0,9 e 0,75 67

Figura 13 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações -0,1; -0,9 e -0,25 68

Figura 14 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações -0,25; 0,25 e 0,5 69

Figura 15 - Ajuste de um modelo não-linear nos parâmetros para descrever a relação entre a correlação dos vetores gerados pela simulação, denominado no gráfico de rho_est (ρ_{est}), e a correlação utilizada na simulação, denominada no gráfico de rho (ρ) 70

Figura 16 - Taxa de Falsos Alertas em relação a média quando $\rho = 0$, $\rho = 0,5$ e $\rho = 0,9$ e $n = 2, 3, 4, 5$ e 10 72

Figura 17 - Taxa de Falsos Alertas em relação a média para $n = 2$, $n = 3$, $n = 4$, $n = 5$, e $n = 10$ e $\rho = 0; 0,5$ e $0,9$ 73

Figura 18 - Taxa de Falsos Alertas em função das correlações para número de armadilhas (n) diferentes 74

Figura 19 - Taxa de Falsos Alertas em função das correlações para médias diferentes 74

Figura 20 - Gráfico dos valores preditos e resíduos estudentizados externamente . . . 76

Figura 21 - Gráfico normal de probabilidades dos resíduos estudentizados do modelo de selecionado 76

Figura 22 - Histograma dos resíduos estudentizados externamente e curva da distribuição normal 77

Figura 23 - Gráfico de caixas (*Box plot*) dos resíduos estudentizados externamente . 77

LISTA DE TABELAS

Tabela 1 - Esquema da análise de variância de regressão 51

Tabela 2 - Estimativas dos coeficientes da regressão e os respectivos erros padrão . 78

1 INTRODUÇÃO

Processos de experimentação que envolvem a avaliação de diversas variáveis respostas nas unidades experimentais são multivariados. Análise multivariada corresponde a um grande número de métodos e técnicas que utilizam simultaneamente as informações de todas as variáveis respostas levantadas na interpretação do conjunto de dados, levando em consideração as correlações existentes entre elas. A disseminação do uso de técnicas multivariadas pode melhorar a qualidade das pesquisas, proporcionar uma economia relativa de tempo e de custo, facilitar a interpretação das estruturas dos dados, diminuindo a perda de informação. (SARTORIO, 2008)

Os dados resultantes de pesquisas no setor agropecuário, como nutrição e alimentação animal, fisiologia animal, melhoramento genético animal ou vegetal, produção e tecnologia de alimentos, dentre outros, são analisados por diversas técnicas estatísticas, sendo a análise de variância univariada e a análise de regressão linear, múltipla e não-linear as mais frequentes. As técnicas de análise multivariada geralmente são mais utilizadas em centros de pesquisa, grandes empresas e no ambiente acadêmico, devido ao desconhecimento dos pesquisadores interessados na pesquisa quantitativa.

Segundo Yahav e Shmueli (2008, 2012) as distribuições multivariadas e principalmente a multinormal são frequentemente utilizadas em problemas complexos na ciência. No entanto, em muitas aplicações, a distribuição normal multivariada não se ajusta aos dados multivariados e existem casos importantes em que os componentes univariados não seguem distribuição normal. Em algumas áreas como hidrologia (YUE; QUARDA; BOBÉE, 2001) e bioestatística (PARK, 2004), por exemplo, algumas variáveis assumem somente valores positivos e seguem distribuições não simétricas como gama, lognormal, inversa Gaussiana, dentre outras.

Na maioria dos casos pode ser difícil especificar completamente a distribuição conjunta do vetor aleatório em estudo. Sendo possível especificar apenas as distribuições marginais dos componentes e a matriz de correlações. A construção de distribuições bivariadas e multivariadas com marginais específicas foi discutida por Farlie (1960), Plackett (1965), Johnson e Tenenbein (1981), Genest e MacKay (1986), Genest (1987) e Koehler e Symanowski (1995), dentre outros, que consideraram somente casos bivariados. Marshall e Olkin (1988), por exemplo, mostraram a construção de distribuições bivariadas por transformada de Laplace, embora a técnica possa ser estendida para casos multivariados.

Koehler e Symanowski (1995), Gross et al. (1998) e Dias, Samaranayaka e Manly (2008) apresentaram um método para construção de distribuições multivariadas. Na primeira etapa eles obtiveram uma distribuição multivariada de variáveis uniformes por transformações das distribuições exponencial e gama. Na segunda etapa, eles substituíram as distribuições uniformes por distribuições acumuladas das marginais desejadas. Este método fornece uma classe de distribuições multivariadas com diversas formas de dependência na qual é pago o preço de um grande grupo de parâmetros dependentes e uma função densidade conjunta complicada.

Outro exemplo de dados que não se aproximam da distribuição normal multivariada são os relativos a contagem e em alguns desses casos, os dados são descritos por uma distribuição de Poisson multivariada. Tais dados aparecem em muitas áreas: na saúde, por exemplo, dados de Poisson multivariada surgem em séries múltiplas de dados de chegadas de pacientes, cada uma relacionada a uma doença diferente ou a hospitais diferentes numa mesma região. Na ciência do marketing e gestão (por exemplo, na gestão de preços e receita) a distribuição de Poisson multivariada pode ser usada para simular compras de produtos complementares ou substitutos. Neste último caso, a correlação entre as compras de produtos de substituição deverá ser negativa.

De acordo com Yahav e Shmueli (2012) a simulação estocástica tem sido usada para avaliar modelos cada vez mais complexos, em estudos de análise de sensibilidade, problemas envolvendo situações reais, dentre outros. A simulação de dados multivariados de Poisson é útil para diferentes funções tais como encontrar as distribuições dos dados multivariados (KARLIS, 2003) e avaliar o desempenho de modelos estatísticos a partir desses dados simulados.

Yahav e Shmueli (2008, 2012), Shin e Pasupathy (2010) apresentam métodos para a geração de variáveis aleatórias correlacionadas com distribuição de Poisson, a partir de transformações da variável aleatória com distribuição normal multivariada, porém esses métodos nem sempre suportam correlações negativas, além da correlação estimada se distanciar da correlação verdadeira quando os valores das médias se aproximam de zero.

2 OBJETIVO

Diante do problema que existe em simular variáveis aleatórias Poisson multivariadas, o presente estudo tem como objetivo:

i) propor um procedimento para gerar dados de distribuições Poisson multivariadas correlacionadas tal que cada vetor aleatório tenha distribuição Poisson univariada e que cada par desses vetores aleatórios, Y_1 e Y_2, tenham média λ_1 e λ_2, e $Corr(Y_1, Y_2) = \rho$, com $\lambda_1, \lambda_2 > 0$ e $\rho \in [-1, 1]$ fixados previamente;

ii) estudar o comportamento das correlações obtidas pelas simulações em função dos valores das correlações fixadas no processo de simulação;

iii) obter as estimativas dos parâmetros de um modelo de regressão não-linear por meio de métodos numéricos quando a correlação fixada e a correlação obtida pela simulação forem estatisticamente diferentes, de forma a entender a relação de viés existente entre essas correlações;

iv) aplicar o procedimento desenvolvido para estimar parâmetros que seguem distribuição Poisson e que estejam correlacionados em modelos estocásticos utilizados na análise da taxa de falsos alertas no controle populacional de percevejos na soja.

3 DESENVOLVIMENTO

3.1 Revisão Bibliográfica

3.1.1 Distribuição Uniforme

A distribuição uniforme representa o mais simples modelo probabilístico contínuo, de fácil operacionalização matemática e ainda com grande aplicabilidade em estatística. Essa aplicabilidade se dá basicamente por duas razões:

1. vários fenômenos aleatórios são modelados pela distribuição uniforme;

2. as simulações estocásticas utilizam a distribuição uniforme para gerar amostras provenientes de modelos probabilísticos mais sofisticados.

Utilizar-se-á a distribuição uniforme no processo de geração de variáveis Poisson correlacionadas e essa característica se deve ao teorema da probabilidade integral que afirma que toda função de distribuição acumulada é, por sua vez, uma variável aleatória com distribuição uniforme de parâmetros 0 e 1, denotada por $U(0,1)$.

A função densidade de probabilidade da uniforme contínua com parâmetros α e β reais, denotada por $U(\alpha,\beta)$ é expressa por:

$$f(x) = \frac{1}{\beta - \alpha} \qquad \beta > \alpha \text{ e } \alpha < x < \beta. \tag{1}$$

A média e a variância da distribuição uniforme são, respectivamente:

$$\mu = \frac{\alpha + \beta}{2} \text{ e } \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

3.1.2 Distribuição Normal Univariada

A distribuição normal é a distribuição de probabilidade mais importante na estatística. As razões disso são muitas e algumas delas são enumeradas a seguir:

1. a grande maioria das técnicas de análise empregadas na estatística são baseadas na distribuição normal;

2. inúmeros são os fenômenos aleatórios cujos comportamentos podem ser descritos precisamente ou de forma aproximada pelo modelo probabilístico normal;

3. a distribuição normal é a forma limitante de muitas outras distribuições de probabilidades, como consequência do teorema central do limite;

4. muitas estatísticas apresentam distribuições assintóticas normais, ou seja, a distribuição da estatística se aproxima da normal à medida que o tamanho da amostra cresce;

5. a álgebra e os cálculos associados à distribuição normal já estão amplamente estudados e definidos.

Uma variável aleatória X com distribuição normal, média μ e variância σ^2 tem função densidade de probabilidade expressa por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]}, \qquad -\infty < \mu < \infty, \quad \sigma^2 > 0, \quad -\infty < x < \infty. \tag{2}$$

Diz-se que X tem distribuição $N(\mu, \sigma^2)$ ou simplesmente que $X \sim N(\mu, \sigma^2)$. A padronização $Z = \frac{X-\mu}{\sigma}$ faz com que essa nova variável aleatória Z tenha média 0 e variância 1, dessa forma:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{\left[\frac{-z^2}{2}\right]}, \qquad -\infty < z < \infty. \tag{3}$$

Para obter a variável aleatória X a partir de Z utiliza-se a expressão $X = \sigma Z + \mu$, em que $E(X) = \mu$ e $Var(X) = \sigma^2$. Esta transformação é utilizada de forma similar no caso multivariado.

3.1.3 Distribuição Poisson

A distribuição Poisson ocorre geralmente quando se deseja contar o número de eventos de certo tipo, que ocorrem em um intervalo de tempo, de superfície ou de volume. A variável Y que segue distribuição Poisson recebeu esse nome em homenagem ao matemático e físico francês S. D. Poisson (1781-1840) e pode assumir infinitos valores inteiros positivos ($y = 0, 1, 2, 3, \ldots$). Uma outra forma de ocorrência de variáveis Poisson surge da aproximação da binomial quando o tamanho da amostra n é considerado "grande" e a probabilidade de sucesso p é considerada "pequena". As definições formais de "grande" para n e "pequeno" para p são difíceis de serem apresentadas, no entanto, de acordo com Ferreira (2009), a aproximação da Poisson para as probabilidades de Binomial é considerada adequada para fins práticos quando $n \geqslant 50$ e $p \leq 0,10$.

Uma variável aleatória Y que segue distribuição Poisson com média e variância iguais a λ, ou simplesmente, $Y \sim Pois(\lambda)$, tem função de probabilidade:

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad y = 0, 1, 2, 3, \ldots \tag{4}$$

3.1.4 Dados Multivariados

Define-se uma observação multivariada como sendo uma coleção de medidas de k variáveis feitas numa mesma unidade experimental ou parcela. A representação desses dados é feita com a notação y_{ij} para indicar um valor particular da i-ésima unidade experimental e da j-ésima variável mensurada. Consequentemente n medidas de k variáveis podem ser arranjadas numa matriz retangular $\boldsymbol{Y}$, com n linhas e k colunas:

$$\boldsymbol{Y}_{n\times k} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1k} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2k} \\ \vdots & \ddots & & & \\ y_{n1} & y_{n2} & y_{n3} & \cdots & y_{nk} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^{'} \\ \boldsymbol{y}_2^{'} \\ \vdots \\ \boldsymbol{y}_n^{'} \end{bmatrix} = [\boldsymbol{y}_1 \ \boldsymbol{y}_2 \ \ldots \boldsymbol{y}_k], \tag{5}$$

em que $\boldsymbol{y}_i^{'}$, $i = 1, ..., n$, é um vetor linha $(1 \times k)$ de observações das k variáveis na parcela; e $\boldsymbol{y}_j$, $j = 1, ..., k$, é um vetor $(n \times 1)$ com valores da variável j nas n parcelas.

Desta forma, cada parcela i é representada no espaço k dimensional por um ponto, em que suas coordenadas são dadas por $\boldsymbol{y}_i^{'} = (y_{i1}, y_{i2}, \ldots, y_{ik})$ para todo $i = 1, 2, \ldots, n$.

O tipo de dados em estatística multivariada também é de grande importância, pois pode ajudar a definir a técnica a ser aplicada na sua análise. De acordo com Hair Jr. et al. (2006), os dados podem ser de dois tipos:

i) Dados métricos: Também chamados de dados quantitativos, dados intervalares ou dados proporcionais, essas medidas identificam ou descrevem indivíduos ou objetos não apenas na posse de um atributo, mas também pela quantia ou grau em que o indivíduo pode ser caracterizado pelo atributo. Por exemplo: a idade ou o peso de um bovino de corte.

ii) Dados não-métricos: Também chamados de dados qualitativos, são atributos, características ou propriedades categóricas que identificam ou descrevem um indivíduo ou objeto. Diferem dos dados métricos no sentido de indicarem a presença de um

atributo, mas não a sua quantia. São também conhecidos como dados nominais ou ordinais. Por exemplo: cor da pelagem, raça etc.

3.1.5 Distribuição Normal Multivariada

Obtém-se uma amostra aleatória com densidade normal multivariada a partir de vetor aleatório $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_k)'$, em que cada $Z_i \sim N(0,1)$ e Z_i, Z_j $(i \neq j)$ são independentes. Dessa maneira $E(\boldsymbol{Z}) = \mathbf{0}$ e $Cov(\boldsymbol{Z}) = \boldsymbol{I}$, em que $\mathbf{0}$ é um vetor $(k \times 1)$ de zeros e $\boldsymbol{I}$ é a matriz identidade k dimensional. Quando as variáveis aleatórias são independentes a função densidade conjunta é o produto das densidades individuais. Dessa maneira,

$$f(\boldsymbol{z}) = f_1(z_1)f_2(z_2)\dots f_k(z_k) = \frac{1}{\sqrt{2\pi}}e^{\left[\frac{-z_1^2}{2}\right]}\frac{1}{\sqrt{2\pi}}e^{\left[\frac{-z_2^2}{2}\right]}\dots\frac{1}{\sqrt{2\pi}}e^{\left[\frac{-z_k^2}{2}\right]} = \frac{1}{(\sqrt{2\pi})^k}e^{-\boldsymbol{z}'\boldsymbol{z}/2}.$$

Diz-se que $\boldsymbol{Z}$ tem distribuição normal multivariada com vetor de média $\mathbf{0}$ e matriz de covariâncias $\boldsymbol{I}$, ou simplesmente $\boldsymbol{Z} \sim N_k(\mathbf{0}, \boldsymbol{I})$. Análogo ao caso univariado, para obter a densidade normal multivariada com vetor de média $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, realiza-se a transformação

$$\boldsymbol{X} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}, \tag{6}$$

em que $\boldsymbol{\Sigma}^{1/2}$ é uma matriz (simétrica) potência meio. O vetor de esperanças e a matriz de covariâncias do vetor aleatório $\boldsymbol{X}$ são

$$E(\boldsymbol{X}) = E(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{1/2}E(\boldsymbol{Z}) + E(\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{1/2}\mathbf{0} + \boldsymbol{\mu} = \boldsymbol{\mu},$$

$$Cov(\boldsymbol{X}) = Cov(\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} + \boldsymbol{\mu}) = \boldsymbol{\Sigma}^{1/2}Cov(\boldsymbol{Z})\left(\boldsymbol{\Sigma}^{1/2}\right)' = \boldsymbol{\Sigma}^{1/2}\boldsymbol{I}\left(\boldsymbol{\Sigma}^{1/2}\right)' = \boldsymbol{\Sigma}.$$

A partir de algumas operações matemáticas, de acordo com Mingoti (2005), tem-se que a função densidade de probabilidade da distribuição normal multivariada é dada por:

$$f(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^k|\boldsymbol{\Sigma}|^{1/2}}e^{-[(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]/2} \tag{7}$$

em que

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2k} \\ \vdots & \ddots & & & \\ \sigma_{k1} & \sigma_{k2} & \sigma_{k3} & \cdots & \sigma_{kk} \end{bmatrix}$$

e por meio da matriz de covariâncias $\boldsymbol{\Sigma}$ é possível obter a matriz de correlações

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{D}^{-1/2} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2k} \\ \vdots & \ddots & & & \\ \rho_{k1} & \rho_{k2} & \rho_{k3} & \cdots & 1 \end{bmatrix}$$

em que $\boldsymbol{D}$ é a matriz diagonal:

$$\boldsymbol{D} = \begin{bmatrix} \sigma_{11} & 0 & 0 & \cdots & 0 \\ 0 & \sigma_{22} & 0 & \cdots & 0 \\ \vdots & \ddots & & & \\ 0 & 0 & \cdots & 0 & \sigma_{kk} \end{bmatrix}.$$

Neste caso, diz-se que $\boldsymbol{X}$ tem distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$, ou simplesmente $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Distribuições Marginais

Todas as distribuições marginais de $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ são normais. Explicitamente, seja $\boldsymbol{X}$ um vetor particionado como:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ e } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

em que as matrizes $\boldsymbol{\Sigma}_{11}$, $\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\Sigma}_{21}$ e $\boldsymbol{\Sigma}_{22}$ possuem dimensões $l \times l$, $l \times (k-l)$, $(k-l) \times l$ e $(k-l) \times (k-l)$ respectivamente. Dessa forma, a distribuição de $\boldsymbol{X}_1$ é $N_l(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ e a distribuição de $\boldsymbol{X}_2$ é $N_{k-l}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, sendo $\boldsymbol{\Sigma}_{12}$ a matriz de covariâncias entre os vetores $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$.

Distribuições Condicionais

A distribuição condicional de $\boldsymbol{X}_1$ dado que $\boldsymbol{X}_2 = \boldsymbol{x}_2$ é

$$N_l\left[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right], \tag{8}$$

em que $\boldsymbol{\Sigma}_{22}$ é não singular. Dessa maneira, a média condicional $E(\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2)$ é uma função linear de $\boldsymbol{x}_2$ enquanto que a covariância da distribuição condicional não depende de $\boldsymbol{x}_2$.

Formas Linear e Quadrática

Seja o vetor aleatório $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{A}$ uma matriz $m \times k$ e $\boldsymbol{a}$ um vetor $m \times 1$. A Distribuição de $\boldsymbol{AX} + \boldsymbol{a}$ é

$$N_m\left(\boldsymbol{A\mu} + \boldsymbol{a}, \boldsymbol{A\Sigma A}'\right). \tag{9}$$

De acordo com Johnson (1987), $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ segue distribuição normal $N_k(\boldsymbol{0}, \boldsymbol{I})$, em que $\boldsymbol{\Sigma}^{-1/2}$ é uma matriz tal que $\left(\boldsymbol{\Sigma}^{-1/2}\right)\left(\boldsymbol{\Sigma}^{-1/2}\right) = \boldsymbol{\Sigma}^{-1}$. Neste caso,

$$\boldsymbol{Z}'\boldsymbol{Z} = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{10}$$

segue distribuição $\chi^2_{(k)}$, se $\boldsymbol{\Sigma}$ é não singular. Este resultado pode ser verificado a partir da função densidade de probabilidade (7), uma vez que $\boldsymbol{x}$ aparece somente em $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Em geral, densidades desse tipo pertencem à classe que possui contornos elípticos, e estas são apresentadas nas Figuras 1, 2, 3 e 4.

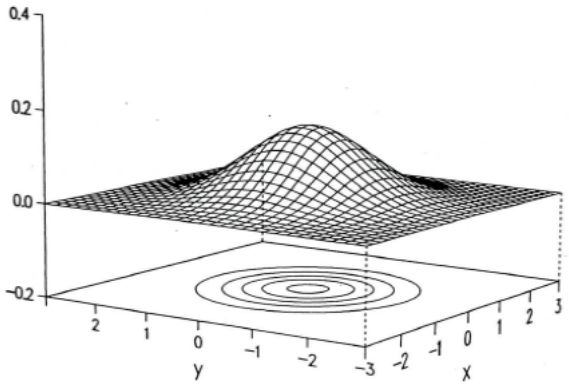

Figura 1 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0$

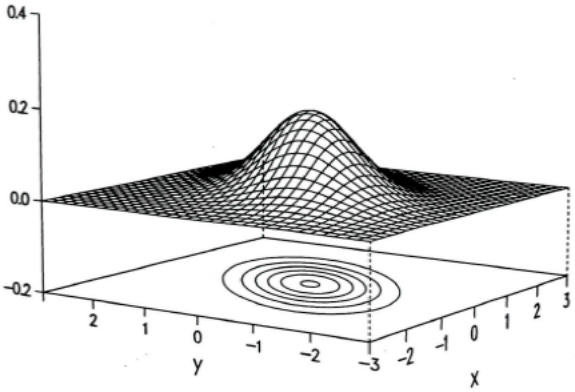

Figura 2 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0,5$

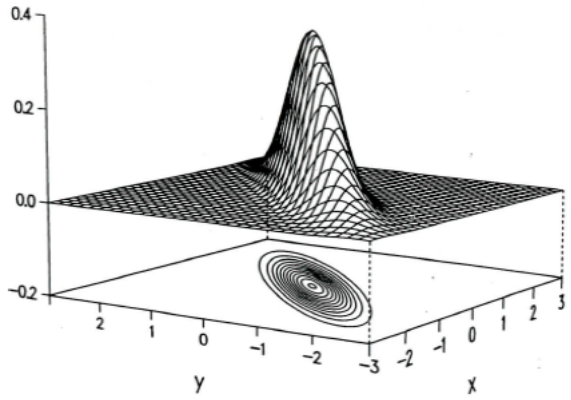

Figura 3 - Distribuição normal bivariada com coeficiente de correlação $\rho = 0,9$

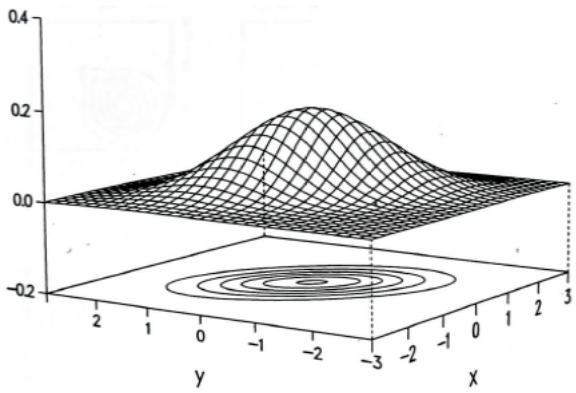

Figura 4 - Distribuição normal bivariada com coeficiente de correlação $\rho = -0,5$

Densidade dos Contornos

Para $\boldsymbol{\Sigma}$ não singular, a função densidade f tem um valor constante no lugar geométrico dos pontos definidos por $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = C$. Este conjunto de pontos corresponde a um elipsóide, cuja probabilidade pode ser deduzida por uma distribuição $\chi^2_{(k)}$. Alguns gráficos de densidades conjuntas para a distribuição normal bivariada com densidades marginais normal padrão são apresentados nas Figuras 1, 2, 3, 4, com a correlação correspondente ρ indicada.

3.1.6 Geração de variáveis aleatórias correlacionadas

3.1.6.1 Geração de variáveis aleatórias de uma distribuição normal multivariada

A distribuição multivariada de probabilidades mais comum é a distribuição normal multivariada k dimensional (ou k variada) com vetor de médias $\boldsymbol{\mu}$ e matriz de

covariâncias $\boldsymbol{\Sigma}$ (denotada por $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), e que segundo Ripley (1987) existem duas maneiras de gerar uma amostra da distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de dispersão $\boldsymbol{\Sigma}$.

Uma delas é supor inicialmente que $\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{S}'$ para alguma matriz $\boldsymbol{S}$ de dimensão $k \times k$. Sejam $Z_1, \ldots, Z_k$ amostras aleatórias obtidas da distribuição normal padrão e assim $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{S}\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, desde que $E\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right] = \boldsymbol{S}E(\boldsymbol{Z}\boldsymbol{Z}')\boldsymbol{S}' = \boldsymbol{S}\boldsymbol{S}'$. A partir da decomposição de Cholesky de $\boldsymbol{\Sigma}$, tem-se que $\boldsymbol{L}$ é a matriz triangular inferior única e que $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{\Sigma}$ (NASH, 1979) e assim $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{Z}$ pode ser formada diretamente.

Uma outra maneira de gerar amostras $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ é gerar uma amostra X_1, então gerar X_2 condicional a X_1, e assim por diante. Seja A_l a submatriz superior de ordem $l \times l$ de $\boldsymbol{\Sigma}$, e $\boldsymbol{a} = (\boldsymbol{\Sigma_{il}})_{1 \leq i \leq l-1}$. Então a distribuição de X_l dado $\boldsymbol{W_l} = (X_1, \ldots, X_{l-1})'$ é $N(\boldsymbol{a}' A_{l-1}^{-1} \boldsymbol{W_l}, \boldsymbol{\Sigma_{ll}} - \boldsymbol{a}' A_{l-1}^{-1} \boldsymbol{a})$. Os vetores $\boldsymbol{a}' A_{l-1}^{-1}$ e os desvios padrão $\sqrt{\boldsymbol{\Sigma_{ll}} - \boldsymbol{a}' A_{l-1}^{-1} \boldsymbol{a}}$ podem ser pré-computados; A_l^{-1} pode ser encontrada de A_{l-1}^{-1} sem inversão completa.

Os métodos da distribuição condicional e de Cholesky diferem apenas em:

$$X_l = \alpha_l Z_l + \beta_{1l} Z_1 + \cdots + \beta_{(l-1)l} Z_{l-1}, \quad 1 \leq l \leq k$$

e para o método de Cholesky

$$X_l = \gamma_l Z_l + \delta_{1l} X_1 + \cdots + \delta_{l\ 1.l} X_l$$

para o método da distribuição condicional.

Estudos comparativos entre esses métodos estão em Scheur e Stoller (1962), Barr e Slezak (1972), Hurst e Knop (1972), dentre outros e o método de Cholesky é preferido devido a sua simplicidade e disponibilidade nos principais pacotes computacionais.

Herndvlgyi (1998) introduziu o seguinte algoritmo para gerar um vetor normal k dimensional, baseado no Teorema do Limite Central:

(a) Gerar $\boldsymbol{Z}$ tal que $\boldsymbol{Z} = 12\boldsymbol{U}(0,1) - 6$, em que $\boldsymbol{U}(0,1)$ denota um vetor da distribuição uniforme com parâmetros 0 e 1. De acordo com o Teorema do Limite Central, a distribuição de $\boldsymbol{Z}$ se aproxima da distribuição $N_k(\boldsymbol{0}, \boldsymbol{I})$.

(b) Seja $\boldsymbol{\Omega}$ uma matriz $k \times k$ em que as colunas são autovetores de $\boldsymbol{\Sigma}$, e $\boldsymbol{\Delta}$ é a matriz diagonal composta pelos autovalores de $\boldsymbol{\Sigma}$. Se $\boldsymbol{Q} = \boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{\Omega}$. Então, $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{Z} + \boldsymbol{\mu} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Em geral, os softwares utilizados em estatística computacional têm a função para gerar dados normais multivariados, como é o caso do software R utilizado nas simulações deste trabalho.

3.1.6.2 Geração de variáveis aleatórias uniformes correlacionadas

A geração de k variáveis uniformes correlacionadas pode ser feita a partir de uma distribuição normal multivariada k dimensional $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, em que para cada variável X_i, $(i = 1, 2, \ldots, k)$ aplica-se a função distribuição normal marginal $F(X_i)$, isto é, $U_i = F(X_i)$. $F(X)$ é calculado a partir da expressão:

$$F(X) = \frac{1}{2} ERF\left(\frac{x}{\sqrt{2}}\right) + \frac{1}{2}$$

em que

$$ERF(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

denominada Função Erro Aleatório (Error Random Function).

Dessa maneira, se X_i e X_j, $(i, j = 1, 2, \ldots, k)$ são variáveis aleatórias normais com correlação ρ_{X_i,X_j}, então as variáveis aleatórias uniformes U_i e U_j são correlacionadas com coeficiente de correlação (Wilks, 1962):

$$\rho_{U_i,U_j} = \frac{6}{\pi} arcsen\left(\frac{\rho_{X_i,X_j}}{2}\right) \tag{11}$$

Embora não seja aparente a partir da equação (11), os valores de ρ_{X_i,X_j} e ρ_{U_i,U_j} são muito semelhantes, pois variando a correlação ao longo do intervalo $[-1, 1]$, de acordo com Dias, Samaranayaka e Manly (2008), a maior diferença absoluta entre ρ_{X_i,X_j} e ρ_{U_i,U_j} foi em torno de 0,02. No mais, invertendo a equação (11), tem-se que a correlação entre X_i e X_j é igual a $\rho_{X_i,X_j} = 2sen\left(\frac{\pi \rho_{U_i,U_j}}{6}\right)$.

3.1.6.3 Geração de variáveis aleatórias Poisson correlacionadas

A distribuição de Poisson k variada é caracterizada por um vetor de médias $\boldsymbol{\Lambda}$ e matriz de covariâncias $\boldsymbol{\Sigma}_{Pois}$ que tem os elementos da diagonal iguais aos elementos de $\boldsymbol{\Lambda}$. A simulação de dados com distribuição Poisson multivariada tem sido muito abordada na literatura, principalmente no caso bivariado. Costuma-se usar o termo "Poisson multivariada" para qualquer extensão da distribuição Poisson univariada, em que os resultados marginais são Poisson univariadas. Em outras palavras, o mesmo termo é usado para

descrever diferentes distribuições multivariadas, que têm em comum a propriedade de suas marginais serem Poisson univariadas. A correlação entre duas variáveis nem sempre pode assumir qualquer valor no intervalo $[-1, 1]$. Diante disso, antes de enunciar os métodos de geração de variáveis aleatórias com distribuição Poisson, explanar-se-á sobre os possíveis valores que a correlação pode assumir.

Estrutura da região viável

Para descrever os possíveis valores que a correlação entre duas variáveis aleatórias com distribuição Poisson, médias λ_1 e λ_2, respectivamente, pode assumir, e consequentemente ilustrar a matrizes de correlações possíveis, Shin e Pasupathy (2010) tomaram como exemplo duas variáveis aleatórias Poisson Y_1 e Y_2 tendo respectivamente médias $\lambda_1 = 0,5$ e $\lambda_2 = 0,5$ e mostraram que a maior correlação alcançável entre Y_1 e Y_2 foi 1, e a menor correlação possível entre Y_1 e Y_2 foi $-0,5$. Portanto, qualquer matriz de correlações

$$\boldsymbol{R}_{Pois} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

em que os possíveis valores de r estão no intervalo $[-0,5; 1]$ é uma matriz de correlações para o vetor aleatório $\boldsymbol{Y} = (Y_1, Y_2)'$. Não existe uma matriz de correlações $\boldsymbol{R}_{Pois}$ se r estiver no intervalo $[-1; -0,5)$.

De modo geral, como mostrado por Whitt (1976), se variáveis aleatórias X_1 e X_2 têm funções de distribuição acumulada $F(x)$ e $G(x)$, respectivamente e U é uma variável aleatória distribuída uniformemente entre 0 e 1, então $Corr(F^{-1}(U), G^{-1}(U))$ é a correlação máxima possível e $Corr(F^{-1}(U), G^{-1}(1-U))$ é a correlação mínima possível entre X_1 e X_2 respectivamente. Portanto o intervalo de correlações possíveis entre X_1 e X_2 é $[Corr(F^{-1}(U), G^{-1}(1-U)), Corr(F^{-1}(U), G^{-1}(U))]$.

As proposições a seguir caracterizam o comportamento limitante da maior e da menor correlação alcançável entre duas variáveis aleatórias Poisson Y_1 e Y_2 com médias λ_1 e λ_2. Assume-se sem perda de generalidade que $\lambda_1 > \lambda_2$ e denota-se a maior e a menor correlação alcançável entre Y_1 e Y_2 por $\rho^+(\lambda_1, \lambda_2)$ e $\rho^-(\lambda_1, \lambda_2)$, respectivamente e $\kappa = \lambda_2/\lambda_1$.

Proposição 1: As funções $\rho^+(\lambda_1, \kappa\lambda_1)$ e $\rho^-(\lambda_1, \kappa\lambda_1)$ são contínuas para $\lambda_1 \in (0, \infty)$ e $\kappa \in (0, 1]$.

Proposição 2: Para um κ fixo,

$$\lim_{\lambda_1 \to 0} \rho^+(\lambda_1, \kappa\lambda_1) = \sqrt{\kappa}; \qquad \lim_{\lambda_1 \to 0} \rho^-(\lambda_1, \kappa\lambda) = 0.$$

Proposição 3: Para um κ fixo,

$$\lim_{\lambda_1 \to \infty} \rho^+(\lambda_1, \kappa\lambda_1) = 1; \qquad \lim_{\lambda_1 \to \infty} \rho^-(\lambda_1, \kappa\lambda_1) = -1.$$

Proposição 4: Para um λ_1 fixo,

$$\lim_{\kappa \to 0} \rho^+(\lambda_1, \kappa\lambda_1) = 0; \qquad \lim_{\kappa \to 0} \rho^-(\lambda_1, \kappa\lambda_1) = 0.$$

Proposição 5: Para um λ_1 fixo,

$$\lim_{\kappa \to 1} \rho^+(\lambda_1, \kappa\lambda_1) = \rho^+(\lambda_1, \lambda_1); \qquad \lim_{\kappa \to 1} \rho^-(\lambda_1, \kappa\lambda_1) = \rho^-(\lambda_1, \lambda_1).$$

Tradicionalmente, o problema de gerar vetores aleatórios de Poisson bivariada é abordado usando um dos dois métodos: (i) redução trivariada (TR) ou (ii) o "normal-to-anything" (Norta).

Redução Trivariada (TR).

É um procedimento bem conhecido em que três variáveis aleatórias Poisson independentes são combinadas apropriadamente para formar duas variáveis aleatórias Poisson correlacionadas (MARDIA, 1970). Especificamente, para gerar variáveis aleatórias Poisson Y_1, Y_2 com os respectivos parâmetros λ_1, λ_2 e correlação $\rho_{Y_1,Y_2} > 0$ a partir do método TR, primeiramente geram-se três variáveis aleatórias Poisson independentes Y_3, Y_4 e Y_{34}, com parâmetros λ_3, λ_4 e λ_{34}, respectivamente. As variáveis aleatórias geradas são então combinadas como $Y_1 = Y_3 + Y_{34}$, $Y_2 = Y_4 + Y_{34}$ para obter Y_1 e Y_2. Como a soma de variáveis aleatórias Poisson independentes é uma variável aleatória Poisson, dessa forma, Y_1, Y_2 são Poisson com parâmetros $\lambda_3 + \lambda_{34}$ e $\lambda_4 + \lambda_{34}$, respectivamente. Os parâmetros λ_3, λ_4 e λ_{34} são escolhidos para obter as médias λ_1, λ_2 e a correlação ρ_{Y_1,Y_2} desejada, a

partir da resolução do sistema:

$$\begin{cases} \lambda_1 = \lambda_3 + \lambda_{34}, \\ \lambda_2 = \lambda_4 + \lambda_{34}, \\ \rho_{Y_1,Y_2} = \dfrac{\lambda_{34}}{\sqrt{(\lambda_3 + \lambda_{34})(\lambda_4 + \lambda_{34})}} \end{cases} \tag{12}$$

Resolvendo o sistema, tem-se:

$$\lambda_{34} = \rho_{Y_1,Y_2}\sqrt{\lambda_1\lambda_2}, \qquad \lambda_3 = \lambda_1 - \lambda_{34} \quad \text{e} \quad \lambda_4 = \lambda_2 - \lambda_{34}.$$

De acordo com Krummenauer (1998, 1999) e Yahav e Shmueli (2012) , duas das mais importantes desvantagens que muitas vezes tornam o TR não utilizável são:

D.1. TR não pode ser usado quando a correlação desejada ρ_{Y_1,Y_2} é negativa.

D.2. Mesmo quando a correlação desejada é positiva, vetores aleatórios obtidos por meio do TR podem não ser capazes de atingir a correlação de interesse, embora as distribuições marginais específicas sejam alcançadas.

A primeira desvantagem é evidente pois a covariância $Cov(Y_1, Y_2) = Var(Y_{34}) = \lambda_{34} > 0$. A segunda desvantagem é notada pelo fato da solução do sistema de equações (12) implicar que λ_1, λ_2 e ρ_{Y_1,Y_2} devem satisfazer $\lambda_1 \geq \rho_{Y_1,Y_2}\sqrt{\lambda_1\lambda_2}$ e $\lambda_2 \geq \rho_{Y_1,Y_2}\sqrt{\lambda_1\lambda_2}$, ou equivalentemente, $\rho_{Y_1,Y_2} \leq \sqrt{\kappa}$ em que $\kappa = Min(\lambda_1, \lambda_2)/Max(\lambda_1, \lambda_2)$. Como λ_3 e λ_4 não são negativos, implica que o TR não pode ser usado para gerar os vetores com as distribuições marginais e correlação desejadas. Isso aponta para um problema bastante sério em TR: quando a discrepância entre as médias λ_3 e λ_4 aumenta, a amplitude do intervalo de correlações que pode ser obtido por TR diminui. De acordo com Shin e Pasupathy (2010), considerando $\lambda_3 = 0,9$ e $\lambda_4 = 9$, as correlações possíveis obtidas por TR estão no intervalo $(0; 0,316)$. Os valores possíveis de ρ_{Y_1,Y_2} estão no intervalo $(-0,8733; 0,9187)$, para todas as combinações de λ_3 e λ_4.

Um dos métodos mais recentes é uma extensão do procedimento TR, proposto por Shin e Pasupathy (2007, 2010) e denominado:

Extensão do Redução Trivariada (TREx).

Este método é uma adaptação do algoritmo TR para a geração de vetores aleatórios bivariados Poisson e que permite correlação negativa, considerando que o objetivo seja gerar o vetor aleatório $\boldsymbol{Y} = (Y_1, Y_2)$ tal que Y_1 tenha distribuição Poisson com

média λ_1, e Y_2 tenha distribuição Poisson com média λ_2 e correlação $Corr(Y_1, Y_2) = \rho$, em que $\lambda_1, \lambda_2 > 0$ e $\rho \in [-1, 1]$ são dados. Essa mesma idéia se estende para gerar distribuições Poisson k dimensionais, com $k > 2$, matriz de covariâncias $\mathbf{\Sigma}_{Pois}$ e vetor de médias $\mathbf{\Lambda}$.

A suposição que $\rho \in [-1, 1]$ cria algumas possibilidades de correlações que são inviáveis, isto é, $\rho > \rho^+(\lambda_1, \kappa\lambda_1)$ ou $\rho < \rho^-(\lambda_1, \kappa\lambda_1)$. Este problema de inviabilidade não é uma complicação porque é automaticamente detectado no fim de uma das etapas do método denominada de pré-processamento. Em outras palavras, no algoritmo proposto, nada de especial precisa ser feito para verificar se há problema de inviabilidade.

Denota-se $F_\lambda^{-1}(y) = inf\{p : F_\lambda(p) > y\}$, em que $F_\lambda(p)$ é a função acumulada da distribuição Poisson com média λ e inf é o ínfimo, ou seja, $inf\{p : F_\lambda(p) > y\}$ é o menor valor de p tal que $F_\lambda(p) > y$. Seja U uma variável aleatória uniformemente distribuída entre 0 e 1. Então o algoritmo proposto tem a seguinte forma:

$$\begin{aligned} Y_1 &= \Upsilon_1 + F_{\lambda^*}^{-1}(U), \quad Y_2 = \Upsilon_2 + F_{\kappa\lambda^*}^{-1}(U) \quad \text{se } \rho > 0; \\ Y_1 &= \Upsilon_1 + F_{\lambda^*}^{-1}(U), \quad Y_2 = \Upsilon_2 + F_{\kappa\lambda^*}^{-1}(1-U) \quad \text{se } \rho < 0. \end{aligned} \tag{13}$$

em que $\Upsilon_1 = F_{\lambda_1-\lambda^*}^{-1}(U_1)$, $\Upsilon_2 = F_{\lambda_2-\kappa\lambda^*}^{-1}(U_2)$, λ^* é um valor obtido na etapa de pré-processamento e U_1 e U_2 são distribuições uniformes $U(0, 1)$ independentes.

Deve-se prestar atenção para três aspectos das operações propostas:

1. Quando $\rho > 0$, isto é, quando a correlação entre Y_1 e Y_2 é positiva, usam-se os números aleatórios gerados pela distribuição uniforme U, como em TR. Quando $\rho < 0$, usam-se variáveis opostas, os seja, usa-se a variável aleatória uniforme U e $1 - U$ para induzir a correlação negativa entre Y_1 e Y_2;

2. Nota-se que para ambos os casos, $\rho > 0$ e $\rho < 0$, ao contrário da TR, não há variáveis aleatória comuns. Em vez disso, variáveis aleatórias com correlação induzida são obtidas por meio da inversão das duas diferentes funções acumuladas da distribuição Poisson. As médias dessas funções acumuladas Poisson estão na mesma proporção que λ_1 e λ_2;

3. O valor de λ^* precisa ser determinado como parte do pré-processamento de modo que as variáveis aleatórias resultantes Y_1 e Y_2 atinjam as médias e a correlação desejada.

As operações envolvidas no TREx estão no algoritmo a seguir, proposto por Shin e Pasupathy (2010) . A etapa 7 do algoritmo consiste de um outro algoritmo, também proposto por Shin e Pasupathy (2010), para obter λ^* (algoritmo 2).

Algoritmo 1

Requer: $\lambda_1 > 0$, $\lambda_2 > 0$, $\rho \in (-1, 1)$

1: **se** $\rho = 0$ **então**

2: Gerar $U_1 \sim U(0,1)$, $U_2 \sim U(0,1)$, independentemente

3: $Y_1 \leftarrow F^{-1}_{\lambda_1}(U_1)$

4: $Y_2 \leftarrow F^{-1}_{\lambda_2}(U_2)$

5: **retorna** (Y_1, Y_2)

6: **fim do se**

7: Resolver para λ^* { Etapa do pré-processamento}

8: Gerar $U_1 \sim U(0,1)$, $U_2 \sim U(0,1)$, $U_3 \sim U(0,1)$, independentemente

9: $\Upsilon_1 \leftarrow F^{-1}_{\lambda_1 - \lambda^*}(U_1)$

10: $\Upsilon_2 \leftarrow F^{-1}_{\lambda_2 - \kappa\lambda^*}(U_2)$

11: $\Upsilon_{12} \leftarrow F^{-1}_{\lambda^*}(U_3)$

12: **se** $\rho > 0$ **então**

13: $\Upsilon'_{12} \leftarrow F^{-1}_{\kappa\lambda^*}(U_3)$

14: **caso contrário**

15: $\Upsilon'_{12} \leftarrow F^{-1}_{\kappa\lambda^*}(1 - U_3)$

16: **fim do se**

17: $Y_1 \leftarrow \Upsilon_1 + \Upsilon_{12}$

18: $Y_2 \leftarrow \Upsilon_2 + \Upsilon'_{12}$

19: **retorna** (Y_1, Y_2)

O procedimento TREx aborda a desvantagem D.1 que existe no procedimento TR. O que não é evidente é o fato do TREx abordar a desvantagem D.2. Para verificar, considere a expressão (13) para $\rho > 0$. A partir da construção, as variáveis aleatórias Υ_1, Υ_2, $F^{-1}_{\lambda^*}(U)$, $F^{-1}_{\kappa\lambda^*}(U)$ seguem distribuição Poisson com médias $\lambda_1 - \lambda^*$, $\lambda_2 - \kappa\lambda^*$,

λ^* e $\kappa\lambda^*$ respectivamente. Portanto, as variáveis aleatórias Y_1 e Y_2 terão distribuições marginais Poisson, desde que $\lambda_1 - \lambda^*$ e $\lambda_2 - \kappa\lambda^*$, sejam positivas. Isto, no entanto, pode ser assegurado restringindo λ^* no intervalo $[0, \lambda_1]$, lembrando que $\lambda_1 \geq \lambda^*$ e $\kappa \leq 1$. Um argumento semelhante é válido para o caso $\rho < 0$.

Considerando o caso $\rho > 0$ na expressão (13), tem-se: que $\lim_{\lambda^* \to 0} Corr(Y_1, Y_2) = 0$, ou seja, um caso não correlacionado. Da mesma forma, quando $\lim_{\lambda^* \to \lambda_1} \lambda_1 - \lambda^* = 0$, $\lim_{\lambda^* \to \lambda_1} \lambda_2 - \kappa\lambda^* = 0$ e $\lim_{\lambda^* \to \lambda_1} \kappa\lambda^* = \lambda_2$. Essas três implicações conjuntas significam que Υ_1 e Υ_2 desaparecem e $Corr(Y_1, Y_2) \to Corr(F^{-1}_{\lambda_1}(U), F^{-1}_{\kappa\lambda_1}(U)) = \rho^+(\lambda_1, \kappa\lambda_1)$. Além disso, pode-se mostrar que $Corr(Y_1, Y_2)$ é uma função contínua de λ^*. Esses três fatos:

- $\lim_{\lambda^* \to 0} Corr(Y_1, Y_2) = 0$;
- $\lim_{\lambda^* \to \lambda_1} Corr(Y_1, Y_2) = \rho^+(\lambda_1, \kappa\lambda_1)$ e
- $Corr(Y_1, Y_2)$ é uma função contínua de λ^*;

asseguram que a extensão completa da correlação positiva $[0, \rho^+(\lambda_1, \kappa\lambda_1)]$ pode ser alcançada por meio do TREx. Da mesma forma, é possível mostrar que a correlação negativa alcançável se encontra no intervalo $[\rho^-(\lambda_1, \kappa\lambda_1), 0]$.

Os argumentos anteriores serão formalmente expostos por meio de uma proposição, porém antes disso, pode-se alcançar um efeito similar para as correlações possíveis por meio da expressão:

$$\begin{aligned} Y_1 &= \Upsilon_1 + F^{-1}_{\lambda_1^*}(U), \quad & Y_2 &= \Upsilon_2 + F^{-1}_{\lambda_2^*}(U) \quad & \text{se } \rho > 0; \\ Y_1 &= \Upsilon_1 + F^{-1}_{\lambda_1^*}(U), \quad & Y_2 &= \Upsilon_2 + F^{-1}_{\lambda_2^*}(1-U) \quad & \text{se } \rho < 0. \end{aligned} \tag{14}$$

ao invés de (13). A expressão (14), contudo, não tem vantagens sobre a expressão (13), pelo menos em duas dimensões e ainda tem a desvantagem de tornar a etapa de pré-processamento uma procura em duas dimensões, em oposição à busca em uma dimensão proporcionada por (13). A demonstração da proposição seguinte é uma simples consequência da proposição 1.

Proposição 6: Sejam Υ_1 e Υ_2 variáveis aleatórias com distribuições Poisson de médias $\lambda_1 - \lambda^*$ e $\lambda_2 - \kappa\lambda^*$, respectivamente, em que $0 < \lambda^* \leqslant \lambda_1$ e $\kappa = \lambda_2/\lambda_1 \leqslant 1$. Seja U uma variável aleatória mutuamente independente de Υ_1 e Υ_2, distribuída uniformemente entre 0 e 1. Então:

i) as funções $Corr\left(\Upsilon_1 + F^{-1}_{\lambda^*}(U), \Upsilon_2 + F^{-1}_{\kappa\lambda^*}(U)\right)$ e $Corr\left(\Upsilon_1 + F^{-1}_{\lambda^*}(U), \Upsilon_2 + F^{-1}_{\kappa\lambda^*}(1-U)\right)$ são contínuas para $\lambda^* \in (0, \lambda_1]$;

ii) $\lim\limits_{\lambda^* \to \lambda_1} \left[Corr(\Upsilon_1 + F^{-1}_{\lambda^*}(U), \Upsilon_2 + F^{-1}_{\kappa\lambda^*}(U))\right] = \rho^+(\lambda_1, \kappa\lambda)$; e

iii) $\lim\limits_{\lambda^* \to \lambda_1} \left[Corr(\Upsilon_1 + F^{-1}_{\lambda^*}(U), \Upsilon_2 + F^{-1}_{\kappa\lambda^*}(1-U))\right] = \rho^-(\lambda_1, \kappa\lambda_1)$;

Etapa de pré-processamento do TREx (encontrando λ^*)

O desafio do procedimento TREx é identificar λ^* tal que a correlação desejada seja obtida. A seguir, descrevem-se os detalhes do método numérico para encontrar λ^*. Da expressão (13), tem-se:

$$Corr(Y_1, Y_2) = \begin{cases} \dfrac{1}{\lambda_1\sqrt{\kappa}}\left(E\left[F^{-1}_{\lambda^*}(U)F^{-1}_{\kappa\lambda^*}(U)\right] - \kappa\lambda^{*2}\right) & \text{se } \rho > 0, \\ \dfrac{1}{\lambda_1\sqrt{\kappa}}\left(E\left[F^{-1}_{\lambda^*}(U)F^{-1}_{\kappa\lambda^*}(1-U)\right] - \kappa\lambda^{*2}\right) & \text{se } \rho < 0. \end{cases} \tag{15}$$

Da expressão 15, identificar λ^* satisfazendo $Corr(Y_1, Y_2) = \rho$ significa, dado λ_1, κ, ρ encontrar λ^* satisfazendo:

$$h(\lambda^*) = \rho\lambda_1\sqrt{\kappa}, \tag{16}$$

em que

$$h(\lambda^*) = \begin{cases} E\left[F^{-1}_{\lambda^*}(U)F^{-1}_{\kappa\lambda^*}(U)\right] - \kappa\lambda^{*2} & \text{se } \rho > 0, \\ E\left[F^{-1}_{\lambda^*}(U)F^{-1}_{\kappa\lambda^*}(1-U)\right] - \kappa\lambda^{*2} & \text{se } \rho < 0 \end{cases} \tag{17}$$

A existência de uma solução para a equação (16) é evidente a partir dos seguintes fatos e do teorema do valor intermediário (BARTLE, 1976):

i) $h(\lambda^*)$ é contínua (Proposição 1),

ii) $\lim\limits_{\lambda^* \to 0} h(\lambda^*) = 0$ (Proposição 2),

iii) $\lim\limits_{\lambda^* \to \infty} h(\lambda^*) = \infty$ se $\rho > 0$ e $-\infty$ se $\rho < 0$ (Proposição 3).

Para encontrar raiz da equação (16) usa-se o método recursivo de Newton em $h(\lambda^*)$, porém, primeiramente é necessário provar que $h(\lambda^*)$ é diferenciável em quase todos os pontos. Isto é provado por meio da proposição 7.

Proposição 7: A função de valores reais

$$h(\lambda^*) = \begin{cases} E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(U)\right] - \kappa\lambda^{*2} & \text{se } \rho > 0, \\ \\ E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(1-U)\right] - \kappa\lambda^{*2} & \text{se } \rho < 0 \end{cases} \tag{18}$$

é diferenciável em quase todos os pontos.

Prova: Desde que os produtos $F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(U)$ e $F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(1-U)$ são não decrescente para λ^* (U fixo), as funções $E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(U)\right]$ e $E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(1-U)\right]$ são ambas não decrescentes em λ^*. Isto implica, portanto, que $E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(U)\right]$ e $E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(1-U)\right]$ são diferenciáveis em quase todos os pontos (ROYDEN, 1988). Portanto, conclui-se que $h(\lambda^*)$ é diferenciável em quase todos os pontos.

A partir do método recursivo de Newton, utilizando um valor inicial λ_0^*, tem-se que:

$$\lambda_{(i+1)}^* = \lambda_{(i)}^* + \frac{1}{h'(\lambda_{(i)}^*)}(\rho\lambda_1\sqrt{\kappa} - h(\lambda_{(i)}^*)). \tag{19}$$

Dessa forma obtêm-se a solução de λ^*, em que $\lambda_{(i)}^*$ é o valor obtido na i-ésima iteração.

Valor inicial

O valor inicial λ_0^* para o método recursivo (19) é obtido a partir da aproximação linear $l(\lambda^*)$ da função $h(\lambda^*)$. Da Proposição 3, para κ fixo, tem-se:

$$\begin{aligned} \lim_{\lambda^*\to\infty} \frac{E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(U)\right] - \kappa\lambda^{*2}}{\lambda^*} &= \sqrt{\kappa}, \\ \lim_{\lambda^*\to\infty} \frac{E\left[F_{\lambda^*}^{-1}(U)F_{\kappa\lambda^*}^{-1}(1-U)\right] - \kappa\lambda^{*2}}{\lambda^*} &= -\sqrt{\kappa}. \end{aligned} \tag{20}$$

Dessa maneira, para um determinado problema, o valor inicial λ_0^* para o método de Newton na expressão (19) é o valor de λ^* na equação $l(\lambda^*) = sign(\rho)\sqrt{\kappa}\lambda^* + c = \rho\lambda_1\sqrt{\kappa}$, ou seja, $\lambda_0^* = (\rho\lambda_1\sqrt{\kappa} - c)/(sign(\rho)\sqrt{\kappa})$, em que $sign()$ é a função sinal, dessa forma, $sign(\rho)$ representa o sinal de ρ e c é uma constante que de acordo com Shin e Pasupathy (2010), $c = 0$ para $\rho > 0$ e $c = \sqrt{\kappa}b_m - \kappa b_m^2$ para $\rho < 0$ em que b_m é a solução da equação $e^{-b_m} + e^{-\kappa b_m} = 1$. O algoritmo apresentado a seguir, proposto por Shin e Pasupathy (2010), é utilizado na etapa de pre-processamento do TREx.

Algoritmo 2

Requer: $\lambda_1 > 0$, $\lambda_2 > 0$, $\rho \in [-1, 1]$, $\epsilon > 0$

1: **se** $|\rho| \leq \epsilon$ **então**

2: **retornar** 0 (isto é, gerar variáveis aleatórias independentes)

3: **fim do se**

4: $\lambda \leftarrow Max(\lambda_1, \lambda_2)$

5: $\kappa \leftarrow Min(\lambda_1, \lambda_2)/\lambda$

6: $s \leftarrow sign(\rho)\sqrt{k}$

7: **se** $\rho < 0$ **então**

8: Obter b_m com erro menor que ϵ

9: **se** $\rho\sqrt{\kappa}\lambda \geq -\kappa b_m^2$ **então**

10: **retornar** $\sqrt{-\rho\lambda/\sqrt{\kappa}}$

11: **fim do se**

12: $c \leftarrow \sqrt{\kappa}b_m - \kappa b_m^2$

13: **caso contrário**

14: $c \leftarrow 0$

15: **fim do se**

16: $\lambda^* = (\rho\sqrt{\kappa}\lambda - c)/s$ (valor inicial)

17: $\hat{\rho} \leftarrow 0$

18: **enquanto** $|\hat{\rho} - \rho| > \epsilon$ **fazer**

19: $$\begin{aligned} \rho^* &= \rho^+(\lambda^*, \kappa\lambda^*) \text{ se } \rho > 0; \\ &= \rho^-(\lambda^*, \kappa\lambda^*) \text{ se } \rho < 0; \\ \rho_1^* &= \tfrac{d\rho^*}{d\lambda}. \end{aligned}$$

20: $\hat{\rho} = \rho^*\lambda^*/\lambda$

21: $h'^{-1} \leftarrow \rho_1^*\sqrt{\kappa}\lambda^* + \rho^*\sqrt{\kappa}$

22: $\lambda^* \leftarrow \lambda^* + h'^{-1}(\rho\sqrt{\kappa}\lambda - \hat{\rho}\sqrt{\kappa}\lambda^*)$

23: **fim do enquanto**

Procedimento NORTA.

NORTA é o método mais geral e popular usado para gerar vetores aleatórios correlacionados (SHIN; PASUPATHY, 2010). O princípio geral em Norta é primeiro gerar variáveis aleatórias com distribuições normais que tenham médias e correlações desejadas para depois transformá-las e assim obter as variáveis aleatórias de interesse. Especificamente, para gerar um vetor aleatório $\boldsymbol{X} = (X_1, X_2, ..., X_k)$ tendo marginais $(F_1, F_2, ..., F_k)$

e matriz de correlações $\boldsymbol{R}_{k\times k}$, usa-se o seguinte procedimento:

1. Gerar um vetor aleatório de uma normal padrão multivariada $\boldsymbol{Z} = (Z_1, Z_2, ..., Z_k)'$, em que $Z_i, i = 1, 2, ..., k$, são variáveis aleatórias da distribuição normal padrão e o vetor $\boldsymbol{Z}$ tem matriz de correlações $\boldsymbol{R}^*_{k\times k}$.

2. Obter $\boldsymbol{X} = (X_1, X_2, ..., X_k)'$ por meio de $X_i = F_i^{-1}(\Phi(Z_i))$ em que $F_i^{-1}(u) = inf\{u : F(x) \geq u\}$ e Φ é função distribuição da normal padrão.

A fase de pré-processamento em Norta envolve a identificação da matriz de correlações $\boldsymbol{R}^*_{k\times k}$, para ser usada na etapa 1 do procedimento Norta, de modo que a matriz de correlações do vetor retornado $\boldsymbol{X}$ na etapa 2 seja $\boldsymbol{R}_{k\times k}$. Para o contexto atual, considerando $k = 2$ e distribuições marginais Poisson com parâmetros λ_1 e λ_2, a fase de pré-processamento em Norta consiste em encontrar um escalar ρ^* tal que $Corr(Z_1, Z_2) = \rho^*$ e

$$Corr(F_1^{-1}(\Phi(Z_1)), F_2^{-1}(\Phi(Z_2))) = \frac{E\left[F_1^{-1}(\Phi(Z_1))F_2^{-1}(\Phi(Z_2))\right] - \lambda_1\lambda_2}{\sqrt{\lambda_1\lambda_2}} = \rho. \qquad (21)$$

O procedimento Norta é muito geral e funciona particularmente bem quando o suporte das marginais necessárias é contínuo, segundo Chen (2001) e Cario e Nelson (1997, 1998). No entanto, quando o suporte de uma ou mais das variáveis aleatórias $X_i = F_i^{-1}(\Phi(Z_i))$, é enumerável, o problema de encontrar a raiz da equação (21) passa a não ser trivial e difícil de resolver de forma eficiente. Segundo Avramidis, Channouf e L'Ecuyer (2009), a principal dificuldade reside na eficiência e precisão para calcular a função $g(\rho^*) = E[F_1^{-1}(\Phi(Z_1))F_2^{-1}(\Phi(Z_2))]$, a qual, no caso enumerável, acaba por ser um somatório duplo infinito começando pela probabilidade da cauda bivariada normal.

Outros métodos para simular dados provenientes de uma distribuição Poisson Multivariada são encontrados na literatura, como o descrito por Minhajuddin, Harris e Schucany (2004), que é baseado na mistura das distribuições Binomial Negativa e Gama. Primeiro, um valor x é gerado de uma distribuição Binomial Negativa com parâmetros r e $\pi = \frac{\lambda}{\lambda+\theta}$. Então, condicional a x, um conjunto de k variáveis Gamma independentes são geradas $(X_1, ..., X_k)$. A soma da distribuição conjunta de x e $X_1, ..., X_k$ tem uma distribuição marginal Gama com parâmetros r e λ, sendo r o parâmetro de forma e λ, o de escala. A correlação entre X_i e X_j $(i \neq j)$ é $\frac{\theta}{\lambda+\theta}$, ou seja, $\theta = \lambda\rho_{\text{gama}}/(1 - \rho_{\text{gama}})$. Nesse método existem duas inconveniências principais: requer que as correlações entre os

pares de variáveis sejam idênticas ($\rho_{ij} = \rho$, para todo $i(i \neq j)$), e não suporta correlação negativa.

Karlis (2003) salienta que o principal obstáculo no uso dos métodos de simulação multivariada, incluindo os mencionados anteriormente, é a complexidade no cálculo da função de probabilidade conjunta e menciona que os somatórios necessários podem ser computacionalmente exaustivos, principalmente quando o número de dimensões k é elevado.

Hahn e Shapiro (1967), descrevem outro método para gerar variáveis aleatórias com distribuição de Poisson correlacionadas. Primeiramente produz-se variáveis aleatórias distribuídas uniformemente no intervalo $(0, 1)$, ou seja $U(0, 1)$, com a correlação desejada. Essas variáveis aleatórias são transformadas a partir da função

$$\sum_{i=1}^{k+1} -\frac{1}{\lambda} \ln(1 - U_i)$$

obtendo assim um valor aleatório y_i para cada distribuição de Poisson Y_i. Em que $y_i = k$ é o menor inteiro tal que o somatório anterior seja maior que 1.

3.1.7 Teste de Aderência

Para comparar se uma determinada distribuição se ajusta aos dados gerados, no caso a Poisson, utilizou-se o teste qui-quadrado denominado de teste de aderência. Segundo Freund e Simon (2000) e Downing e Clarck (2002), para testar se as diferenças entre as frequências observadas f de uma determinada ocorrência com as frequências esperadas f^* podem ser atribuídas ao acaso, utiliza-se a estatística qui-quadrado:

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - f_i^*)^2}{f_i^*} \tag{22}$$

Se o valor de χ^2 calculado exceder o valor χ^2_α, rejeita-se a hipótese nula sobre a qual se baseiam as frequências esperadas, ao nível α de significância e conclui-se que a distribuição testada não se ajusta aos dados. O número de graus de liberdade é $n - m - 1$, em que n é o número de termos $\frac{(f_i - f_i^*)^2}{f_i^*}$ adicionados na fórmula de χ^2 e m é o número de parâmetros de distribuição de probabilidade que devem ser estimados com base nos dados amostrais.

3.1.8 Simulação

Usou-se o software R (R DEVELOPMENT CORE TEAM, 2010) para gerar os dados, a partir de função *mvrnorm* da biblioteca *MASS* para gerar vetores da Normal multivariada, a função *pnorm* para obter o vetor de probabilidades da Normal e a função *dpois* para obter o vetor de probabilidade da distribuição de Poisson. O código de programação está nos anexos. Como a relação entre correlação verdadeira (utilizada na distribuição normal) e a correlação obtida após o processo de simulação tem um formato sigmóide, um modelo de regressão não linear nos parâmetros foi utilizado para descrever essa tendência.

3.1.9 Regressão Não-Linear

Segundo Draper e Smith (1998), Bates e Watts (1988), Ratkowsky (1983) dentre outros, a definição para que um modelo seja designado não-linear é que pelo menos uma das derivadas parciais da variável dependente com relação a algum parâmetro presente no modelo, dependa pelo menos de um outro parâmetro. Desse modo, os modelos de regressão foram classificados por Draper e Smith (1998) como modelos de regressão intrinsecamente lineares e não-lineares.

Gallant (1987) afirma que uma das situações mais comuns em análise estatística é aquela de dados observados que consistem em respostas univariadas y_i, conhecidas sendo dependentes da contribuição x_i. Essa situação pode ser representada por equações de regressão:

$$y_i = f(x_i, \boldsymbol{\theta}) + e_i, \tag{23}$$

em que $i = 1, 2, ..., \mathrm{n}$, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)$ é um vetor k dimensional de parâmetros desconhecidos e e_i é o erro aditivo, representando fenômenos não observáveis ou o erro experimental. Os erros são assumidos independentes e identicamente distribuídos com média zero e variância desconhecida σ^2.

A sequência de valores x_i também é chamada de ajuste experimental, de forma que o vetor de parâmetros e o ajuste experimental juntos determinam o valor da função modelo $f(x_i, \boldsymbol{\theta})$ que, sobre as suposições de que $E[e_i] = 0$, é o valor esperado de y_i condicional a x_i e $\boldsymbol{\theta}$ (BATES; WATTS, 1980):

$$E[y_i | x_i, \boldsymbol{\theta}] = f(x_i, \boldsymbol{\theta})$$

A função modelo é freqüentemente chamada de modelo ou modelo matemático, mas existe uma distinção que deve ser feita, tendo em vista que funções modelos diferentes podem ser associadas com o mesmo modelo. A função modelo é associada com a parametrização particular escolhida para o modelo, então, um conjunto diferente de parâmetros pode produzir uma função modelo diferente. O modelo é a relação que dá suporte entre a resposta e o ajuste experimental gerado pela função modelo e é independente da parametrização.

Estimação dos parâmetros

Método dos mínimos quadrados

Define-se a soma de quadrados dos erros para o modelo não-linear fornecido pelos dados como (DRAPER; SMITH, 1998):

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i - f(x_i, \boldsymbol{\theta})\}^2 \tag{24}$$

Desde que y_i e x_i sejam observações fixas, a soma dos quadrados dos erros é uma função de $\boldsymbol{\theta}$. A estimativa de mínimos quadrados $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ é o valor que minimiza $S(\boldsymbol{\theta})$. Pode ser mostrado que a estimativa de mínimos quadrados de $\boldsymbol{\theta}$ é a mesma estimativa obtida pelo método da máxima verossimilhança quando os erros são independentes e seguem distribuição normal de média zero e variância σ^2 (DRAPER; SMITH, 1998). Para encontrar o estimador de mínimos quadrados $\hat{\boldsymbol{\theta}}$ é necessário diferenciar a equação (24) com relação a $\boldsymbol{\theta}$, gerando k equações normais que precisam ser resolvidas para $\hat{\boldsymbol{\theta}}$. As equações normais podem ser tomadas na forma:

$$\sum_{i=1}^{n} \{y_i - f(x_i, \boldsymbol{\theta})\} \left[\frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_t}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \tag{25}$$

para $t = 1, 2, ..., k$ em que a quantidade denotada pelos colchetes são as derivadas parciais de $f(x_i, \boldsymbol{\theta})$ em relação a θ_t com todos os $\theta's$ substituídos pelos correspondentes $\hat{\theta}'s$, que têm o mesmo subscrito.

Considerando para cada ajuste experimental x_i a resposta condicional esperada:

$$\eta_i(\boldsymbol{\theta}) = E[y_i|\boldsymbol{\theta}] = f(x_i, \boldsymbol{\theta}) \tag{26}$$

e os vetores $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \eta_2(\boldsymbol{\theta}), ..., \eta_n(\boldsymbol{\theta}))'$, $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$ e $\boldsymbol{e} = (e_1, e_2, ..., e_n)'$, então pode-se escrever o modelo em forma vetorial, como:

$$\boldsymbol{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{e} \tag{27}$$

e a função soma de quadrados dos erros como:

$$S(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 \tag{28}$$

em que as $\|\|$ representa a norma do vetor. Assim tem-se uma interpretação geométrica do quadrado da distância entre o vetor $\boldsymbol{y}$ e o vetor $\boldsymbol{\eta}(\boldsymbol{\theta})$ em um espaço amostral n dimensional. À medida que $\boldsymbol{\theta}$ é modificado, o vetor $\boldsymbol{\eta}(\boldsymbol{\theta})$ descreve uma superfície k dimensional, chamada de "o local da solução" nesse espaço amostral (BOX; LUCAS, 1959). Em geral, $\boldsymbol{y}$ não será um ponto no local da solução e a estimativa de mínimos quadrados $\hat{\boldsymbol{\theta}}$ é um ponto no local da solução próximo de $\boldsymbol{y}$.

O método dos mínimos quadrados é utilizado na estimação dos parâmetros em modelos não-lineares da mesma maneira que em modelos lineares. Porém, para o sistema de equação normais não-linear não existe uma solução explícita, a qual deve ser obtida por meio de processos iterativos por aproximação em série de Taylor, usando um software estatístico (GALLANT, 1987).

A maioria dos algoritmos utilizados para obter as estimativas de mínimos quadrados $\hat{\boldsymbol{\theta}}$ e a maioria dos métodos inferenciais para modelos não-lineares são baseados em métodos iterativos que consideram uma aproximação linear local para o modelo (BATES; WATTS, 1980).

Método da máxima verossimilhança

A aplicação do método de mínimos quadrados não exige qualquer pressuposição além da formulação do modelo. A incorporação da pressuposição de normalidade aos erros ($e_i \sim N(0, \sigma^2)$) vai implicar também que

$$y_i \sim N(f(x_i, \boldsymbol{\theta}), \sigma^2) \tag{29}$$

ou que as observações (y_i) têm distribuição normal.

A distribuição conhecida dos erros (e das observações) nos permite estimar os parâmetros pelo método da máxima verossimilhança, em que a função densidade de

probabilidade é dada por:

$$h_{Y_i}(y_i, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(y_i - f(x_i, \boldsymbol{\theta}))^2\right)}. \tag{30}$$

Se $Y_1, Y_2, ..., Y_n$ é uma amostra aleatória independente, então a função de verossimilhança do modelo não-linear pode ser obtido a partir da seguinte expressão

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} h_{Y_i}(y_i, \boldsymbol{\theta}) \tag{31}$$

e consequentemente a função de verossimilhança do modelo não-linear como função de $\boldsymbol{\theta}$, é expressa na seguinte forma

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} S(\boldsymbol{\theta}; \boldsymbol{y})\right) \tag{32}$$

em que $S(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} (y_i - f(x_i, \boldsymbol{\theta}))^2$. Assim a estimativa obtida pelo método da máxima verossimilhança (EMV) de $\boldsymbol{\theta}$ é o $\hat{\boldsymbol{\theta}}$ que maximiza $L(\boldsymbol{\theta}; \boldsymbol{y})$.

A EMV $\hat{\boldsymbol{\theta}}$ é obtida por um processo interativo e esta estimativa é consistente e tem distribuição assintótica normal k variada de média $\boldsymbol{\theta}$ e estrutura de variância-covariância $\boldsymbol{K}^{-1} = \sigma^2(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}$, em que $\tilde{\boldsymbol{X}}$ é a matriz Jacobiana da transformação de $f(x_i, \boldsymbol{\theta})$ em $\boldsymbol{\theta}$ (CORDEIRO; LIMA NETO, 2004). Analogamente à regressão linear, a estimativa mais usual para σ^2 é dada por $s^2 = S(\hat{\boldsymbol{\theta}}; \boldsymbol{y})/(n-k)$, sendo que $S(\hat{\boldsymbol{\theta}}; \boldsymbol{y})$ é a soma de quadrados dos resíduos do modelo ajustado. Logo, um intervalo de $100(1-\alpha)\%$ para θ_j, será formado pelos limites

$$\hat{\theta}_j \pm t_{\alpha/2}(-\hat{k}^{jj})^{1/2} \tag{33}$$

em que $t_{\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma distribuição t - Student com $(n-k)$ graus de liberdade e $-\hat{k}^{jj}$ é a estimativa do elemento (j,j) da matriz $\boldsymbol{K}^{-1}$ definida anteriormente.

Uma região de aproximadamente $100(1-\alpha)\%$ de confiança para $\boldsymbol{\theta}$ foi proposto por Beale (1960), e é formada pelos contornos de $S(\boldsymbol{\theta}; \boldsymbol{y})$ tais que

$$S(\boldsymbol{\theta}; \boldsymbol{y}) = S(\hat{\boldsymbol{\theta}}; \boldsymbol{y}) \left\{1 + \frac{k}{n-k} F_{n,n-k}(\alpha)\right\}. \tag{34}$$

Em particular, se $L(\boldsymbol{\theta}; \boldsymbol{y})$ for aproximadamente quadrática, a região de confiança acima é bem aproximada por

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq s^2 k F_{k,n-k}(\alpha), \tag{35}$$

em que $F_{k,n-k}(\alpha)$ é o quantil $(1-\alpha)$ de uma distribuição F e a matriz $\tilde{\boldsymbol{X}}$ é avaliada em $\hat{\boldsymbol{\theta}}$. Essa última expressão é uma adaptação da região de confiança para os parâmetros da regressão linear.

Para testar a hipótese $H_0 : \hat{\boldsymbol{\theta}} \in \Theta_0$, em que Θ_0 é um subconjunto do espaço paramétrico, utiliza-se usualmente a estatística da razão de verossimilhanças, dada por

$$-2\log\lambda = n\log\left\{S(\tilde{\boldsymbol{\theta}};\boldsymbol{y}) - S(\hat{\boldsymbol{\theta}};\boldsymbol{y})\right\}, \tag{36}$$

em que $S(\tilde{\boldsymbol{\theta}};\boldsymbol{y})$ é a soma dos quadrados dos resíduos para o modelo ajustado em H_0. Sob essa hipótese, a estatística acima tem assintoticamente distribuição χ^2 com $(k-m)$ graus de liberdade, em que $m = dim(\Theta_0)$, ou seja m é o número de parâmetros.

Uma estatística alternativa para testar $H_0 : \hat{\boldsymbol{\theta}} \in \Theta_0$ é dada por

$$F = \frac{n-k}{k-m}\frac{\left\{S(\tilde{\boldsymbol{\theta}};\boldsymbol{y}) - S(\hat{\boldsymbol{\theta}};\boldsymbol{y})\right\}}{S(\hat{\boldsymbol{\theta}};\boldsymbol{y})}, \tag{37}$$

que sob H_0 tem assintoticamente distribuição F com $(k-m)$ e $(n-k)$ graus de liberdade.

Aproximação linear

Na vizinhança de um valor do vetor de parâmetros $\boldsymbol{\theta}^0$ a função modelo é aproximada por:

$$f\left(x_i,\boldsymbol{\theta}\right) \cong f\left(x_i,\boldsymbol{\theta}^0\right) + \sum_{t=1}^{k}\left(\theta_t - \theta_t^0\right)v_t(x_i)$$

em que

$$v_t(x_i) = \left[\frac{\partial f\left(x_i,\boldsymbol{\theta}\right)}{\partial\theta_t}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$$

Reunindo os componentes dentro dos vetores $\boldsymbol{\eta}\left(\boldsymbol{\theta}\right)$, $\boldsymbol{\eta}\left(\boldsymbol{\theta}^0\right)$ e $\boldsymbol{v}_t$, sendo que

$$\begin{aligned}\boldsymbol{v}_t &= \left(v_t(x_1), v_t(x_2), ..., v_t(x_n)\right)' = \\ &= \left[\tfrac{\partial\boldsymbol{\eta}}{\partial\theta_t}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}, \ t = 1, 2, ..., k\end{aligned}$$

têm-se:

$$\boldsymbol{\eta}\left(\boldsymbol{\theta}\right) \cong \boldsymbol{\eta}\left(\boldsymbol{\theta}^0\right) + \sum_{t=1}^{k}\left(\theta_t - \theta_t^0\right)\boldsymbol{v_t}$$

ou

$$\boldsymbol{\eta}\left(\boldsymbol{\theta}\right) \cong \boldsymbol{\eta}\left(\boldsymbol{\theta}^0\right) + \boldsymbol{V}^0\left(\boldsymbol{\theta} - \boldsymbol{\theta}^0\right) \tag{38}$$

em que $\boldsymbol{V}^0$ é a matriz de derivadas avaliada em $\boldsymbol{\theta}^0$ tendo como colunas os vetores $\boldsymbol{v}_t's$ ordenadamente. A matriz $\boldsymbol{V}^0$ também é chamada de a matriz Jacobiana com ordem $n \times k$ de $\boldsymbol{\eta}(\boldsymbol{\theta})$ avaliada em $\boldsymbol{\theta}^0$:

$$\boldsymbol{V}^0 = \begin{bmatrix} \frac{\partial \eta_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial \eta_1(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial \eta_1(\boldsymbol{\theta})}{\partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \eta_n(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial \eta_n(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial \eta_n(\boldsymbol{\theta})}{\partial \theta_k} \end{bmatrix}_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$$

Bates e Watts (1980) afirmam que o efeito dessa aproximação é substituir o local da solução pelo seu plano tangente em $\boldsymbol{\eta}(\boldsymbol{\theta}^0)$, e simultaneamente impor um sistema de coordenadas uniformes nesse plano tangente. Esses dois componentes da aproximação linear em (38) são chamados "pressuposição planar" e "pressuposição de coordenadas uniformes" respectivamente. Tanto a efetividade do algoritmo de mínimos quadrados quanto a validade das inferências feitas com respeito aos parâmetros de um modelo não-linear serão afetadas pela proximidade da linearidade do modelo.

O método de Gauss-Newton

A abordagem sugerida por Gauss foi utilizar uma aproximação linear a função esperança como em (38) para reiterativamente melhorar o $\boldsymbol{\theta}^0$ como valor inicial, estimar $\boldsymbol{\theta}$ e continuar melhorando as estimativas até que a diferença entre as estimativas obtidas na t-ésima e $(t-1)$-ésima iterações seja tão pequena quanto se queira. Segundo Ratkowsky (1983) esse método expande uma série de Taylor sobre $\boldsymbol{\theta}_t$ (o subscrito t indica a t-ésima iteração) como em (38):

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \cong \boldsymbol{\eta}(\boldsymbol{\theta}_t) + \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$

Rearranjando (38) com (28), tem-se:

$$\begin{aligned} S(\boldsymbol{\theta}) &= \|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 = [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})]' [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})] \cong \\ &\cong [\boldsymbol{y} - (\boldsymbol{\eta}(\boldsymbol{\theta}_t) + \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t))]' [\boldsymbol{y} - (\boldsymbol{\eta}(\boldsymbol{\theta}_t) + \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t))] = \\ &= [(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)) - \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)]' [(\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)) - \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)] = \\ &= [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)]' [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)] - 2.[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)]' \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + (\boldsymbol{\theta} - \boldsymbol{\theta}_t)' \boldsymbol{V}_t' \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t) \end{aligned}$$

O vetor gradiente $g(\boldsymbol{\theta}) = \left(\frac{\partial S(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_2}, ..., \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_d}\right)'$, é então:

$$g(\boldsymbol{\theta}) = -2\boldsymbol{V}_t' [\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}_t)] + 2\boldsymbol{V}_t' \boldsymbol{V}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t)$$

Igualando essa expressão a zero, e rearranjando os valores,obtém-se:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \left[\boldsymbol{V}_t^{'}\boldsymbol{V}_t\right]^{-1} \boldsymbol{V}_t^{'} \left[\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{\theta}_t\right)\right] \tag{39}$$

Com valores iniciais para $\boldsymbol{\theta}$ com $t=1$, o processo continua até a convergência, que ocorre quando $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|$ é menor do que algum valor pré-fixado.

Draper e Smith (1998) discutem alguns problemas do método da linearização:

i) Ele pode convergir muito lentamente, isto é, um grande número de iterações pode ser requerido para que a solução estabilize, mesmo que a soma de quadrados $S\left(\boldsymbol{\theta}_t\right)$ decresça de forma constante à medida que t cresce.

ii) Pode oscilar bastante, continuamente mudando de direção e muitas vezes de modo crescente, decrescendo a soma de quadrados. Todavia, a solução pode estabilizar eventualmente;

iii) Pode não convergir sob qualquer condição e inalteradamente divergir.

Observe que o método de Gauss-Newton serve também para encontrar as estimativas em modelo de regressão linear. Tendo em vista que tal modelo pode ser escrito da forma:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{e}$$

em que $\boldsymbol{X}$ é uma matriz de ordem $n \times k$ de variáveis regressoras, cuja primeira coluna pode ser um vetor unitário se a função modelo inclui um termo constante. Para a função descrita anteriormente, a matriz Jacobiana é:

$$\boldsymbol{V}_0 = \boldsymbol{X}$$

Para qualquer valor arbitrário utilizado como estimativa inicial, por exemplo, $\boldsymbol{\theta}_0$, o vetor de estimativas atualizado $\boldsymbol{\theta}_1$ é:

$$\begin{aligned}
\boldsymbol{\theta}_{0+1} &= \boldsymbol{\theta}_0 + \left[\boldsymbol{V}_0^{'}\boldsymbol{V}_0\right]^{-1} \boldsymbol{V}_0^{'} \left[\boldsymbol{y} - \boldsymbol{\eta}\left(\boldsymbol{\theta}_0\right)\right] \\
&= \boldsymbol{\theta}_0 + \left[\boldsymbol{X}^{'}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^{'} \left[\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_0\right] \\
&= \boldsymbol{\theta}_0 + \left[\boldsymbol{X}^{'}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^{'}\boldsymbol{y} - \left[\boldsymbol{X}^{'}\boldsymbol{X}\right]^{-1} \left[\boldsymbol{X}^{'}\boldsymbol{X}\right] \boldsymbol{\theta}_0 \\
&= \boldsymbol{\theta}_0 + \left[\boldsymbol{X}^{'}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^{'}\boldsymbol{y} - \boldsymbol{\theta}_0 \\
&= \left[\boldsymbol{X}^{'}\boldsymbol{X}\right]^{-1} \boldsymbol{X}^{'}\boldsymbol{y}
\end{aligned}$$

Segundo Draper e Smith (1998), a estimativa fica imutável se o processo iterativo é continuado. O lado direito da igualdade é o estimador de mínimos quadrados para um modelo de regressão linear. Assim, o método de Gauss-Newton, para um modelo linear, converge para o estimador de mínimos quadrados em uma única iteração para qualquer valor inicial.

Além do método de Gauss-Newton, existem outros métodos iterativos propostos na literatura para obtenção das estimativas de mínimos quadrados dos parâmetros de um modelo de regressão não-linear como: o método Steepest-Descent ou método Gradiente; o método de Marquardt e o método de Newton-Raphson. Em geral esses métodos podem ser utilizados quando o método de Gauss-Newton apresenta problemas com relação a convergência (BATES; WATTS, 1988).

Gallant (1987) afirma que o sucesso em termos de convergência para $\hat{\boldsymbol{\theta}}$ não é garantido por qualquer método. A experiência indica que a falha no processo de convergência depende da distância do valor inicial à resposta correta e do grau de parametrização da função resposta relativamente ao conjunto de dados utilizado. Quando a convergência não ocorre, deve-se tentar encontrar valores iniciais melhores ou utilizar uma função resposta mais adequada.

Valores iniciais

Os valores iniciais podem ser estimativas preliminares baseadas em qualquer informação disponível. Por exemplo: podem ser valores sugeridos por informação ganha num ajuste similar de equações em um laboratório ou sugerida por experiências anteriores e conhecimentos empíricos do pesquisador.

Esses valores serão utilizados em iterações sucessivas, com o objetivo de serem as mais baixas descrições, ou seja, quanto melhor eles forem, menor será o número de iterações necessárias para se encontrar a convergência. Vários princípios simples, porém úteis para determinar valores iniciais podem ser usados, como sugerido por Bates e Watts (1988) e Draper e Smith (1998):

i) interpretar o comportamento da função esperança em relação aos parâmetros;

ii) interpretar o comportamento das derivadas da função esperança em relação aos parâmetros;

iii) transformar a função esperança para obter um comportamento mais simples, preferencialmente linear. No caso do erro ser aditivo, pode-se desconsiderar a sua existência apenas para a obtenção das estimativas iniciais;

iv) reduzir dimensões substituindo valores para alguns parâmetros ou avaliando a função nos valores do delineamento específico;

3.1.10 Regressão Linear Múltipla

O modelo estatístico para a regressão linear múltipla com k variáveis regressoras é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + e_i \tag{40}$$

em que:

β_j $(j = 0, \ldots, k)$ são os parâmetros do modelo;

X_{ij} $(i = 1, \ldots, n)$ são as variáveis regressoras;

e_i é o erro aleatório.

Utilizando a notação matricial, o modelo de regressão linear múltipla apresenta-se como:

$$\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e} \tag{41}$$

ou seja,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \ddots & & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ e } \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \tag{42}$$

em que:

$\boldsymbol{y}$ é o vetor de observações $(n \times 1)$ de variável aleatória Y;

$\boldsymbol{X}$ é a matriz $(n \times (k+1))$ de variáveis regressoras;

$\boldsymbol{\theta}$ é o vetor $(k \times 1)$ dos coeficientes de regressão;

$\boldsymbol{e}$ é o vetor $(n \times 1)$ de erros aleatórios.

As suposições para o modelo de regressão linear múltipla são:

i) a variável resposta $\boldsymbol{y}$ é função linear das variáveis regressoras X_j, $j = 1, 2, ...k$;

ii) as variáveis X_j são fixas;

iii) $E(e_i) = 0$ e $Var(e_i) = \sigma^2$;

iv) os erros têm distribuição normal;

v) os erros são independentes.

Combinando (iii) e (v) tem-se $Var(\boldsymbol{e}) = E(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{I}\sigma^2$, sendo $\boldsymbol{I}$ uma matriz identidade, de dimensões $n \times n$. Assim, considerando-se, também, (v) tem-se que $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ e $\boldsymbol{y} \sim N(\boldsymbol{X\theta}, \boldsymbol{I}\sigma^2)$, pois $E(\boldsymbol{y}) = \boldsymbol{X\theta}$ e $Var(\boldsymbol{y}) = I\sigma^2$.

3.1.11 Método dos Mínimos Quadrados

O método utilizado para obter as estimativas dos parâmetros do modelo é o método dos mínimos quadrados (MONTGOMERY; PECK, 1992). Considerando o modelo (41), o vetor de erros é dado por:

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X\theta} \tag{43}$$

O método dos mínimos quadrados minimiza o comprimento do vetor de erros ($\boldsymbol{e}$). Usando a norma euclidiana para a obtenção do comprimento de $\boldsymbol{e}$, tem-se:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} e_i^2 = (\boldsymbol{y} - \boldsymbol{X\theta})'(\boldsymbol{y} - \boldsymbol{X\theta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\theta}'\boldsymbol{X}'\boldsymbol{X\theta}. \tag{44}$$

Derivando $S(\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$, obtém-se:

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X\theta}$$

e igualando $\dfrac{\partial S}{\partial \boldsymbol{\theta}}$ a zero, obtém-se:

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{y}$$

denominado de sistema de equações normais, que é consistente e apresenta solução única dada por:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{45}$$

sendo que $\hat{\boldsymbol{\theta}}$ é um estimador não viciado de $\boldsymbol{\theta}$, isto é, $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ e com variância dada por:

$$Var(\hat{\boldsymbol{\theta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2. \tag{46}$$

3.1.12 Teste t e teste F

Utilizando a estatística do teste t, dado por:

$$t = \frac{\hat{\beta}_i - \beta_i}{\sqrt{s\left(\hat{\beta}_i\right)}}, \tag{47}$$

em que $H_0: \ \beta_i = 0$ versus $H_1: \ \beta_i \neq 0, \ i = 1, 2, ..., k$, verifica-se a significância dos efeitos no modelo e a estatística anterior tem distribuição t-Student com $k-1$ graus de liberdade

Para os efeitos significativos realiza-se a análise de variância (Tabela 1) e verifica-se, individualmente, a inclusão das variáveis não significativas pelo teste F.

Tabela 1 - Esquema da análise de variância de regressão

Fonte de Variação	GL	SQ	QM	F
Regressão	w	SQ_{Reg}	$\frac{SQ_{Reg}}{w}$	$\frac{SQ_{Reg}}{QM_{Res}}$
Resíduo	$n-w-1$	SQ_{Res}	$\frac{SQ_{Res}}{(n-w-1)}$	
Total	$n-1$	SQ_{Total}		

em que:

$$SQ_{Reg} = \hat{\boldsymbol{\theta}}_R' \boldsymbol{X}_R' \boldsymbol{y} - C \tag{48}$$

sendo $\hat{\boldsymbol{\theta}}_R$ o vetor, de ordem $w+1$, dos parâmetros referentes aos efeitos significativos no modelo, $w \leq k$; e $\boldsymbol{X}_R$ a matriz das variáveis correspondentes,

$$SQ_{Total} = \sum_{i=1}^{n} y_i^2 - C; \text{ sendo } C = \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \tag{49}$$

e a soma de quadrados dos resíduos é obtida por diferença.

3.1.13 Coeficiente de Correlação Múltiplo

O coeficiente de determinação, denotado por R^2, representa a proporção da variação de Y, explicada pela regressão (DRAPER; SMITH, 1998) e é obtido por:

$$R^2 = \frac{SQ_{Reg}}{SQ_{Total}} = 1 - \frac{SQ_{Res}}{SQ_{Total}}, \ 0 \leq R^2 \leq 1. \tag{50}$$

O R^2 deve ser usado com cuidado para avaliar a qualidade do ajuste (DEMÉTRIO; ZOCCHI, 2006), pois:

i) depende do tamanho da amostra (n), tendendo a crescer quando n diminui;

ii) seu valor pode aumentar com a inclusão de uma nova variável no modelo;

iii) a magnitude de R^2, também, depende da amplitude de variação das variáveis regressoras. Em geral, o valor de R^2 aumenta com a amplitude de variação dos X's e diminui em caso contrário.

Assim, o R^2 deve ser considerado em conjunto com outros diagnósticos do modelo. Com a finalidade de tentar corrigir estes problemas foi definido o coeficiente de determinação ajustado para graus de liberdade, denotado por $\overline{R}^2$. Da equação (50), tem-se:

$$1 - R^2 = 1 - \frac{SQ_{Reg}}{SQ_{Total}} = \frac{SQ_{Res}}{SQ_{Total}}. \tag{51}$$

Substituindo SQ_{Res} e SQ_{Total}, na equação (50), pelos respectivos quadrados médios, obtém-se o coeficiente de determinação ajustado (MONTGOMERY. PECK, 1992), dado por:

$$\overline{R}^2 = 1 - \frac{n-1}{n-p}(1 - R^2) \tag{52}$$

em que p é o número de parâmetros do modelo de regressão.

3.1.14 Métodos de seleção de variáveis

Segundo Demétrio e Zocchi (2006) o objetivo de uma análise de regressão linear múltipla é obter a melhor descrição possível de variável dependente (Y), em função das variáveis regressoras (X). Um número muito grande de variáveis regressoras pode levar a um modelo que explique bem os dados, mas de difícil interpretação. Por outro lado, um número pequeno de variáveis regressoras pode facilitar a interpretação, porém, o ajuste pode não ser adequado. O ideal é um modelo intermediário que produza um bom ajuste aos dados e com o menor número possível de parâmetros. Os critérios utilizados para a seleção de variáveis no modelo são:

1. Coeficiente de determinação ajustado;

2. A estatística C_p (MALLOWS, 1973) dada por:

$$C_p = \frac{SQ_{Res_p}}{s^2} - (n - 2p), \tag{53}$$

em que SQ_{Res_p} é a soma de quadrados do resíduo do modelo, com p parâmetros; s^2 é o quadrado médio do resíduo referente ao modelo ajustado com todas as variáveis regressoras (modelo completo) e n é o número de observações.

Para $k = p - 1$ variáveis selecionadas, valores grandes de C_p indicam modelos com QM_{Res} grandes. Se o modelo possui $C_p > p$ é um indicatico de que o modelo está mal especificado. Por outro lado, se $C_p < p$, é um indicativo de que o modelo está superparametrizado. Uma alternativa é ajustar todos os possíveis modelos e escolher o que possui o menor número de parâmetros e valor da estatística C_p menor ou igual ao do número de parâmetros do modelo.

3. Critério de informação de Akaike (AIC): é definido por

$$AIC = -2 \log L - 2p \tag{54}$$

em que L é a função de verossimilhança e p o número de parâmetros do modelo ajustado.

4. Teste F: a comparação entre dois modelos com p e q parâmetros, sendo $p < q$, pode ser feita utilizando a estatística:

$$F = \frac{SQ_{Res_p} - SQ_{Res_q}}{(q - p)QM_{Res}}, \tag{55}$$

sendo QM_{Res} uma estimativa de σ^2, em geral, obtida a partir do modelo de regressão completo.

Os métodos utilizados para a seleção de variáveis no modelo são:

1. Todas as regressões possíveis: consiste no ajuste de todos os modelos, em que os modelos são separados em grupos, de acordo com o número de variáveis, sendo estão, cada grupo ordenado de acordo com algum critério de seleção de variáveis.

2. Método do passo atrás (*backward*): esse método busca examinar os principais modelos de regressão, com um certo número de variáveis. Consiste em ajustar, inicialmente, o modelo completo e eliminar, uma a uma as variáveis com base na :

i) menor correlação parcial com a resposta Y;

ii) menor diminuição do R^2;

iii) menor diminuição significativa no teste F parcial ou no teste t parcial, de acordo com algum critério de parada.

3. Método do passo a frente (*forward*): esse método supõe que não existem variáveis regressoras no modelo, além do intercepto. Consiste em incluir no modelo a variável com maior coeficiente de correlação simples com a variável resposta e, na sequência, incluir, uma a uma, variáveis com:

i) maior correlação parcial com a resposta Y;

ii) maior aumento do R^2;

iii) maior aumento significativo no teste F parcial.

4. Método do passo a frente passo atrás (*stepwise*): esse método é uma modificação do método do passo a frente, na qual, em cada passo, todas as variáveis adicionadas previamente no modelo são avaliadas por meio da estatística F parcial ou pelo critério de informação de Akaike.

3.1.15 Análise de resíduos

Para o modelo linear e não linear $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e}$ tem-se que o vetor $\boldsymbol{e}$ contém os elementos $e_i's$ que são as diferenças entre os valores observados (y_i) e os valores ajustados ($\hat{y}_i$).

Admite-se que os e_i são independentes e que $e_i \sim N(0, \sigma^2)$ seguindo as suposições do modelo. Quando as suposições são violadas, a análise do modelo pode levar a conclusões duvidosas.

Montgomery; Peck e Vining (2006) definem três tipos de resíduos:

1. Resíduos ordinários: utilizando o método dos mínimos quadrados, para o modelo linear $\boldsymbol{y} = \boldsymbol{X\theta} + \boldsymbol{e}$, tem-se que os resíduos ordinários são obtidos por:

$$\hat{e}_i = y_i - \hat{y}_i$$

Os valores ajustados são obtidos por:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\theta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \tag{56}$$

sendo que $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é chamada de matriz de projeção. Os elementos da matriz $\boldsymbol{H}$ são denotados por h_{ij}, ou seja

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}, \tag{57}$$

e os elementos da diagonal da matriz $\boldsymbol{H}$, denomina-se h_i de *leverage* e fornece a distância da i-ésima observação em relação as demais observações. Quando o modelo de regressão passa pela origem, $0 \leq h_i \leq 1$. A matriz $\boldsymbol{H}$ é simétrica ($\boldsymbol{H} = \boldsymbol{H}'$) e idempotente ($\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$).

Utilizando a matriz $\boldsymbol{H}$, os resíduos podem ser expressos por:

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \tag{58}$$

porém, enquanto os erros $e_i's$ são independentes e de mesma variância σ^2, não se pode dizer o mesmo para os resíduos ordinários, pois $Var(\hat{\boldsymbol{e}}) = Var[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}] = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$ e a variância para o i-ésimo resíduo é dada por: $Var(\hat{e}_i) = (1 - h_{ii})\sigma^2$ e a covariância entre o i-ésimo e o j-ésimo resíduos por: $Cov(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2$. Assim, o uso dos resíduos ordinários não é o mais indicado para a análise dos resíduos, devido a heterogeneidade de variâncias. Foram, então, propostas diferentes padronizações para sanar esse problema.

2. Resíduos estudentizados internamente: sabe-se que o quadrado médio do resíduo (s^2) é uma estimativa para σ^2. Assim, uma estimativa para $Var(\hat{e}_i)$ é dada por:

$$\widehat{Var}(\widehat{e_i}) = (1 - h_{ii})s^2$$

e os resíduos estudentizados internamente são definidos por

$$rei_{(i)} = \frac{\widehat{e_i}}{\sqrt{\widehat{Var}(\hat{e}_i)}} = \frac{y_i - \hat{y}_i}{s\sqrt{1 - h_{ii}}}. \tag{59}$$

Estes resíduos são assim chamados, pois a estimativa da variância residual é obtida considerando todas as observações. Portanto, essa estimativa pode variar, consideravelmente, quando tem-se um valor discrepante. Além disso, o numerador e o

denominador estão correlacionados, isto é $Cov\left[\hat{e}_i, \widehat{Var}(\hat{e}_i)\right] \neq 0$.

3. Resíduos estudentizados externamente: pode-se garantir a independência entre o numerador e o denominador, na padronização dos resíduos, definindo-se os resíduos estudentizados externamente por

$$ree_{(i)} = \frac{\hat{e}_i}{s_{(i)}\sqrt{1-h_{ii}}}, \tag{60}$$

em que $s^2_{(i)}$ é a estimativa de σ^2 livre da influência da observação (i). A vantagem de utilizar os resíduos estudentizados externamente é que, sob normalidade, $ree_{(i)}$ tem distribuição t de *Student* com $(n-p-1)$ graus de liberdade.

Diagnóstico em influência

Em algumas situações, pode ocorrer que um conjunto pequeno de observações cause uma influência muito grande sobre o modelo de regressão ajustado. Nesse caso, é necessário localizar esses pontos influentes a avaliar seu impacto no ajuste do modelo (MONTGOMERY; PECK; VINING, 2006).

Para o devido ajuste do modelo, uma observação pode ser classificada, de acordo com Demétrio e Zocchi (2006), como:

1. Inconsistente: quando tem valor do $ree_{(i)}$ muito diferente da tendência geral das demais. Geralmente, é inconsistente se

$$|ree_{(i)}| \geq t_{\left\{1-\left(\frac{\gamma}{2n}\right);n-p-1\right\}}, \tag{61}$$

   com nível de significância de $100\gamma\%$. Uma observação inconsistente é considerada um ponto discrepante (*outlier*) quando possui medida de *leverage* (h_i) pequeno e resíduo $ree_{(i)}$ grande.

2. Ponto de alavanca: quando possui h_i grande, geralmente $h_i \geq 2p/n$, sendo considerado bom quando consistente e ruim quando inconsistente.

3. Influente: uma observação é considerada influente quando sua omissão no conjunto de dados provoca mudanças substanciais no ajuste do modelo.

Estatística para diagnósticos

1. DFBETAS: Montgomery; Peck (1992) sugerem uma estatística que mede a alteração em $\hat{\boldsymbol{\theta}}$ quando a i-ésima observação é retirada da análise, e que é obtida por:

$$DFBETAS_{j,i} = \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{s_{(i)}\sqrt{C_{jj}}} \tag{62}$$

   em que: C_{jj} é o j-ésimo elemento da diagonal da matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ e $\hat{\theta}_{j(i)}$ é o j-ésimo coeficiente de regressão calculado sem a i-ésima observação.

2. DFFITS: mede a alteração provocada no valor ajustado quando a i-ésima observação é retirada da análise e é definida como:

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}}, \ i = 1, 2, ..., n, \tag{63}$$

   sendo que $\hat{y}_{(i)}$ é o valor ajustado de y_i obtido sem considerar a i-ésima observação.

3. Distância de Cook: é uma medida de afastamento do valor de $\hat{\theta}$ quando a i-ésima observação é retirada. É semelhante ao DFFITS, utilizando como estimativa da variância residual o resíduo estudentizado internamente. É obtida por:

$$D_{(i)} = \frac{h_{ii}}{(1 - h_{ii})}\frac{1}{p}(rei_{(i)})^2. \tag{64}$$

4. COVRATIO: mede a influência da i-ésima observação na precisão da estimativa do vetor $\boldsymbol{\theta}$. Montgomery; Peck e Vining (2006) definem $COVRATIO$ como:

$$COVRATIO_i = \frac{|(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}s^2_{(i)}|}{|(\boldsymbol{X}'\boldsymbol{X})^{-1}s^2|}, \ i = 1, 2, ..., n \tag{65}$$

   Se $COVRATIO_i > 1$, a i-ésima observação melhora a precisão da estimativa, enquanto que se $COVRATIO_i < 1$, a inclusão da i-esima observação degrada a precisão.

   Computacionalmente,

$$COVRATIO_i = \frac{(s^2_{(i)})^p}{(s^2)^p}\left(\frac{1}{1 - h_{ii}}\right), \ i = 1, 2, ..., n \tag{66}$$

   Note-se que $[1/(1 - h_{ii})]$ é a razão entre $|(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}|$ e $|(\boldsymbol{X}'\boldsymbol{X})^{-1}|$, de modo que um ponto de alavanca elevado fará $COVRATIO_i$ elevado.

No programa estatístico R, a i-ésima observação é considerada influente se ao menos em um dos casos ocorrer (DEMÉTRIO, ZOCCHI, 2006):

1. $|DFBETAS_{(i)}| > 1$;
2. $|DFFITS_{(i)}| > 3\sqrt{p/(n-p)}$;
3. $|1 - COVRATIO| > 3p/(n-p)$;
4. $D_{(i)} > F_{\{50\%;p;n-p\}}$;
5. $h_i > 3p/n$,

3.2 Material e Métodos

Os valores utilizados nas matrizes de correlações para as simulações são: $-0,95$; $-0,90$; $-0,75$; $-0,50$; $-0,25$; $-0,10$; 0; $0,10$; $0,25$; $0,50$; $0,75$; $0,90$ e $0,95$, que já foram utilizados por Dias, Samaranayaka e Manly (2008). Já os valores das médias utilizadas nas simulações são: 1, 2, 5, 10, 20 e 40. Considerando os 13 possíveis valores que as correlações podem assumir, existem $13^3 = 2197$ matrizes que são possíveis candidatas a matriz de correlações $\boldsymbol{R}_N$. Dessas 2197 matrizes, verificou-se que 539 são positivas definidas, logo tem-se 539 matrizes candidatas a $\boldsymbol{R}_N$.

Das combinações possíveis para a matriz de correlações, foram realizadas algumas simulações para verificar se o método Norta mantinha a estrutura de correlação original $\boldsymbol{R}_N$, ou seja, se a matriz de correlações gerada $\boldsymbol{R}_{Pois}$ se aproximava de $\boldsymbol{R}_N$. Para cada combinação entre a matriz de correlações e vetor de médias foram realizadas 1000 simulações e o teste de aderência χ^2 foi utilizado para verificar se a distribuição de Poisson desejada se ajustava às variáveis aleatórias geradas. Essas simulações foram divididas em alguns casos, como mostrado a seguir.

Para cada um dos casos, geraram-se 1000 distribuições normais trivariadas, cada uma contendo 500 observações trivariadas, vetor de médias $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ e matriz de correlação $\boldsymbol{R}_N$. A partir de cada uma dessas distribuições normais trivariadas obteve-se uma Poisson trivariada com vetor de médias $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \lambda_3)$ e matriz de correlações $\boldsymbol{R}_{Pois}$. Embora tenha-se optado por trabalhar com a matriz de correlações, era possível trabalhar com a matriz de covariâncias em que para isso, de acordo com Yahav e Shmueli (2012), $\boldsymbol{\Sigma}_N$, $\forall i, j \in p : Cov(X_i, X_j) = \rho_{ij}\sigma_i\sigma_j$, em que ρ_{ij} é a correlação entre X_i e X_j.

Como a relação entre correlação verdadeira (ρ), utilizada na distribuição normal, e a correlação obtida pelo processo de simulação (ρ_{est}) tem um formato sigmóide, um modelo de regressão não-linear nos parâmetros foi utilizado para descrever essa tendência. As estimativas dos parâmetros foram obtidas por meio do método numérico de Gauss-Newton.

Utilizou-se o método de geração de variáveis aleatórias com distribuição Poisson multivariada correlacionada desenvolvido para estimar parâmetros que sigam essa distribuição e que estejam correlacionados num modelo estocástico utilizado no controle populacional de percevejos *Euschistus heros* (Fabricius), que são considerados pragas na cultura da *Glycine max L* (soja).

Utilizou-se a regressão linear múltipla para estimar a taxa de falsos alertas em função da quantidade média esperada de percevejos em cada armadilha, do número de armadilhas e da correlação existente entre os números de insetos capturados nas armadilhas. Para a seleção de variáveis utilizou-se o método passo a frente passo atrás (*stepwise*) e análise do resíduo foi realizada por meio do gráfico dos valores ajustados *versus* resíduos estudentizados externamente, gráfico normal de probabilidade (*Q-Q plot*), teste de normalidade de *Shapiro-Wilk* e gráficos de caixas (*box plot*), todos realizados com os resíduos estudentizados externamente.

As estimativas dos parâmetros foram obtidas pelo método dos mínimos quadrados e o coeficiente de determinação ajustado foi utilizado para verificar o quanto da variabilidade foi explicada pelo modelo.

4 RESULTADOS

4.1 Resultados das Simulações

4.1.1 Médias e correlações iguais

a) Iniciou-se com alguns casos específicos para a geração da distribuição Poisson Multivariada em que as correlações da matriz de correlação eram iguais e os elementos do vetor de médias também iguais. Considerando o vetor de médias e a matriz de correlações:

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

obteve-se, a partir das simulações, o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1,07220 \\ 1,01230 \\ 1,01011 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,40382 & 0,41036 \\ 0,40382 & 1 & 0,40472 \\ 0,41036 & 0,40472 & 1 \end{bmatrix}.$$

A partir do teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 1 se ajustou aos 500 valores simulados para as variáveis aleatórias que representam as marginais da distribuição, em uma das simulações, pois os valores p foram iguais a $0,08$; $0,44$ e $0,85$ (Figura 5). Além disso, a correlação entre os vetores formados pelos valores simulados ficou abaixo da correlação desejada, com uma diferença média de $0,09$ enquanto que as médias obtidas ficaram aproximadamente $0,03$ acima das médias estipuladas.

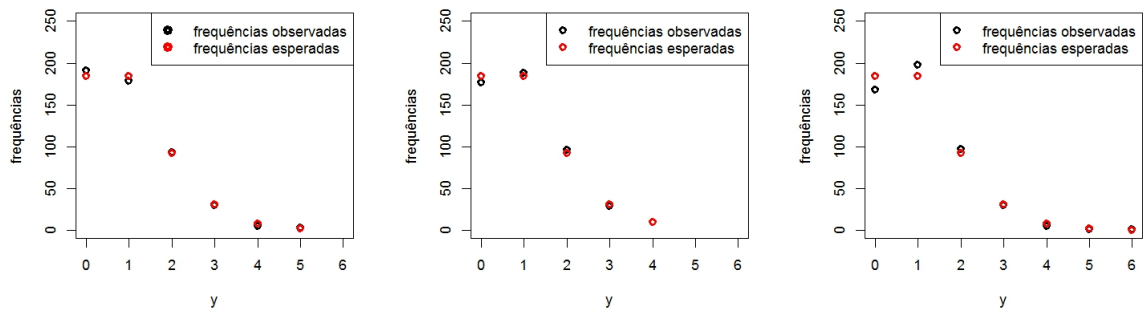

Figura 5 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 1

b) Para o vetor de médias e matriz de correlações:

$$\boldsymbol{\mu} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1,94623 \\ 1,99424 \\ 1,98443 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,41540 & 0,42820 \\ 0,41540 & 1 & 0,41070 \\ 0,42820 & 0,41070 & 1 \end{bmatrix}.$$

A partir do teste de aderência qui-quadrado concluiu-se que a distribuição de Poisson com média 2 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, em uma das simulações, pois os valores p foram iguais a $0,88$; $0,90$ e $0,63$ (Figura 6). No mais, as médias dos valores obtidos nas simulações ficaram aproximadamente $0,03$ abaixo das médias desejadas que eram iguais a 2 e as correlações simuladas ficaram abaixo das correlações desejadas, com uma diferença média de $0,08$

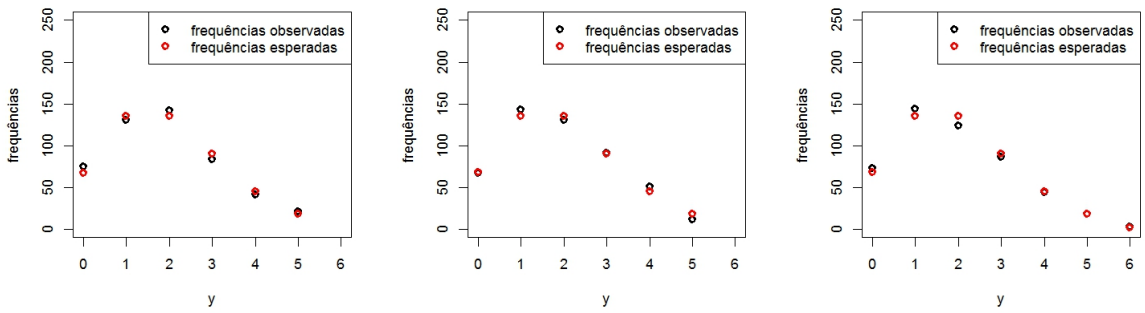

Figura 6 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 2

c) Considerando o vetor de médias e a matriz de correlações utilizados nas simulações:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5,03739 \\ 5,04033 \\ 5,03486 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,46718 & 0,46864 \\ 0,46718 & 1 & 0,46896 \\ 0,46864 & 0,46896 & 1 \end{bmatrix}.$$

Realizou-se o teste de aderência qui-quadrado as três distribuições marginais, em uma das simulações, obtendo os valores p iguais a $0,66$; $0,19$ e $0,69$, com isso concluiu-se que a distribuição poisson com média 5 se ajustou aos valores simulados os quais representam as marginais (Figura 7). A diferença entre as médias dos valores simulados e as médias atribuídas no início da simulação foi de aproximadamente $0,04$ e as correlações simuladas embora tenham ficado abaixo das correlações desejadas, a diferença diminuiu para em média $0,03$.

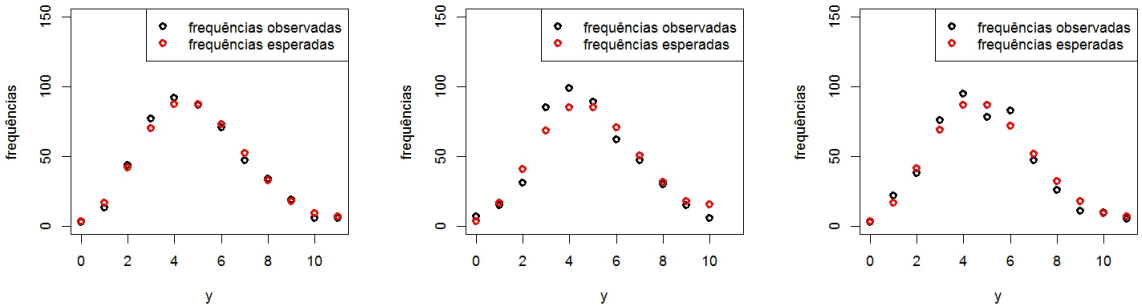

Figura 7 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 5

d) Para o vetor de médias e a matriz de correlações:

$$\boldsymbol{\mu} = \begin{bmatrix} 10 \\ 10 \\ 10 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 10,00316 \\ 9,99952 \\ 10,00363 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,49211 & 0,49514 \\ 0,49211 & 1 & 0,49555 \\ 0,49514 & 0,49555 & 1 \end{bmatrix}$$

Verificou-se, a partir do teste de aderência qui-quadrado (valores p iguais a 0,43; 0,12 e 0,85 respectivamente), que os valores simulados das variáveis aleatórias as

quais representam as marginais seguem distribuição Poisson com média igual a 10 (Figura 8) .

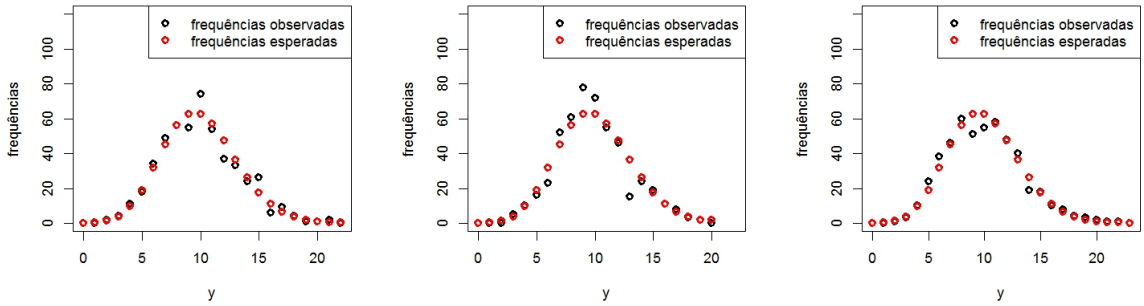

Figura 8 - Frequências observadas e frequências esperadas dos valores simulados das variáveis aleatórias com distribuição Poisson e média 10

e) Para o vetor de médias e matriz de correlações:

$$\boldsymbol{\mu} = \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 20,00745 \\ 20,00855 \\ 19,99935 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,49709 & 0,49525 \\ 0,49709 & 1 & 0,49652 \\ 0,49525 & 0,49652 & 1 \end{bmatrix}.$$

A partir do teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 20 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, com valores p iguais a $0,43$; $0,91$ e $0,91$ respectivamente (Figura 9).

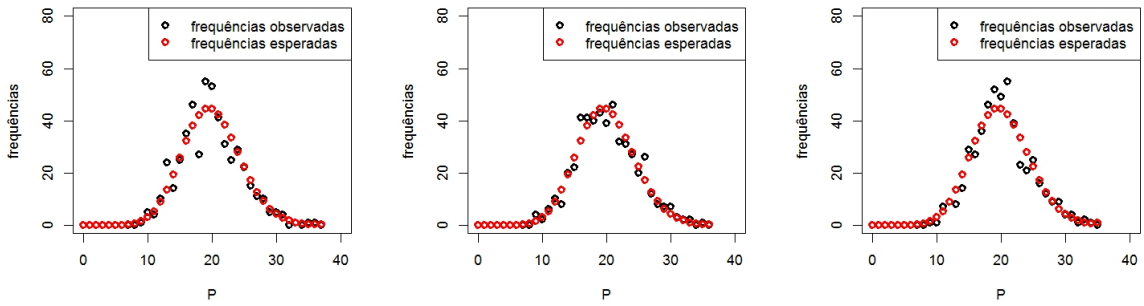

Figura 9 - Frequências observadas e frequências esperadas dos valores simulados das variáveis aleatórias com distribuição Poisson e média 20

f) Para o vetor de médias e matriz de correlações:

$$\boldsymbol{\mu} = \begin{bmatrix} 40 \\ 40 \\ 40 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & 0,5 \\ 0,5 & 1 & 0,5 \\ 0,5 & 0,5 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 40.00099 \\ 40.00120 \\ 40.00768 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,49835 & 0,49889 \\ 0,49835 & 1 & 0,49861 \\ 0,49889 & 0,49861 & 1 \end{bmatrix}$$

A partir do teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 40 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, com valores p iguais a $0,88$; $0,17$ e $0,20$ respectivamente (Figura 10).

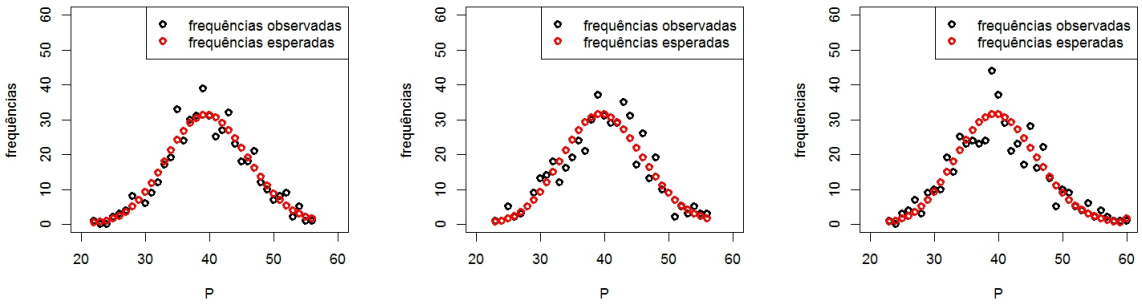

Figura 10 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson e média 40

De acordo com as simulações anteriores, em que se manteve a mesma estrutura de correlação entre as variáveis aleatórias com distribuições marginais normais e variou-se os valores das médias de uma simulação para a outra, observou-se que não existe problema no método de simulação, pois a distribuição Poisson se ajustou, em todos os casos, aos valores simulados das variáveis aleatórias, ao nível de $0,05$ de significância, concluindo-se com isso que as marginais dos dados trivariados simulados seguem distribuição Poisson. Na medida que as médias aumentaram, a diferença entre as correlações desejadas e as correlações obtidas entre os vetores simulados diminuíram, o que era esperado uma vez que quando a média aumenta a distribuição Poisson se aproxima da distribuição normal.

4.1.2 Médias iguais e correlações diferentes

a) Iniciou-se com alguns casos de simulação da distribuição Poisson multivariada com correlações diferentes e os elementos do vetor de médias iguais. Considerando:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,5 & -0,5 \\ 0,5 & 1 & 0,25 \\ -0,5 & 0,25 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5,03592 \\ 5,03440 \\ 5,02953 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,46760 & -0,44917 \\ 0,46760 & 1 & 0,23069 \\ -0,44917 & 0,23069 & 1 \end{bmatrix}$$

A partir do teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 5 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, com valores p iguais a $0,07$; $0,26$ e $0,26$ (Figura 11).

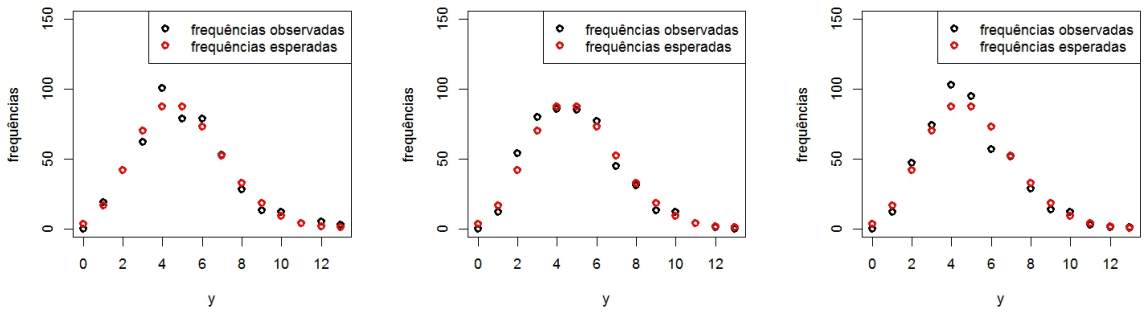

Figura 11 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações 0,5; -0,5 e 0,25

b) Mantendo o vetor de médias e alterando os valores dos elementos na matriz de correlações tem-se:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & 0,9 & -0,9 \\ 0,9 & 1 & -0,75 \\ -0,9 & -0,75 & 1 \end{bmatrix}$$

e a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5,02791 \\ 5,03690 \\ 4,99525 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & 0,86766 & -0,79530 \\ 0,86766 & 1 & -0,66623 \\ -0,79530 & -0,66623 & 1 \end{bmatrix}$$

De acordo com o teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 5 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, com valores p iguais a $0,48$; $0,40$ e $0,40$ (Figura 12).

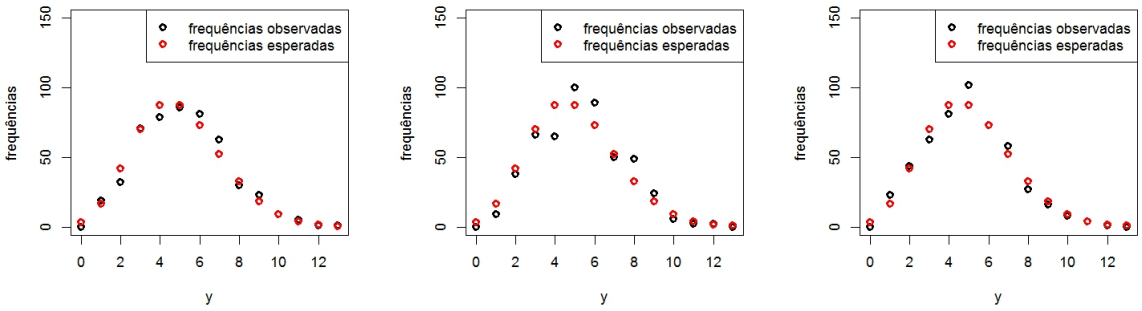

Figura 12 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações 0,9; -0,9 e 0,75

c) Considerando:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} \text{ e } \boldsymbol{R}_{N} = \begin{bmatrix} 1 & -0,1 & 0,9 \\ -0,1 & 1 & -0,25 \\ 0,9 & -0,25 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5,03445 \\ 5,02591 \\ 5,03017 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & -0,08981 & 0,86870 \\ -0,08981 & 1 & -0,22578 \\ 0,86870 & -0,22578 & 1 \end{bmatrix}$$

De acordo com o teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 5 se ajustou aos valores simulados das variáveis aleatórias as quais representam as marginais da distribuição, com valores p iguais a $0,28$; $0,53$ e $0,15$ (Figura 13).

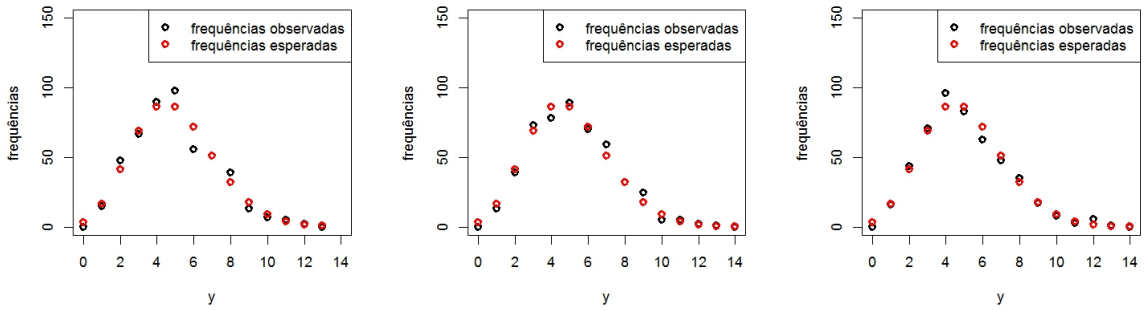

Figura 13 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações -0,1; -0,9 e -0,25

d) Considerando:

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} \text{ e } \boldsymbol{R}_N = \begin{bmatrix} 1 & -0,25 & 0,25 \\ -0,25 & 1 & 0,5 \\ 0,25 & 0,5 & 1 \end{bmatrix}$$

a partir das simulações obteve-se o vetor de médias e a matriz de correlações:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 5,04126 \\ 5,03238 \\ 5,03807 \end{bmatrix} \text{ e } \boldsymbol{R}_{Pois} = \begin{bmatrix} 1 & -0,22311 & 0,23119 \\ -0,22311 & 1 & 0,46862 \\ 0,23119 & 0,46862 & 1 \end{bmatrix}$$

De acordo com o teste de aderência qui-quadrado concluiu-se que a distribuição Poisson com média 5 se ajustou aos valores simulados os quais representam as marginais da distribuição, com valores p iguais a $0,12$; $0,22$ e $0,15$ (Figura 14).

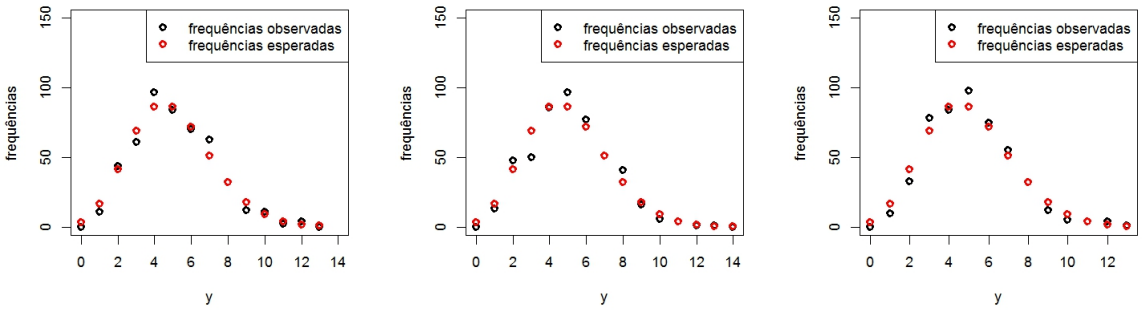

Figura 14 - Frequências observadas e frequências esperadas dos 500 valores simulados das variáveis aleatórias com distribuição Poisson, média 5 e correlações -0,25; 0,25 e 0,5

De acordo com as simulações anteriores, em que se manteve o mesmo vetor de médias com valores iguais a 5 e variou-se os elementos de matriz de correlações entre as variáveis aleatórias com distribuições marginais normais, observou-se que a distribuição Poisson se ajustou aos valores simulados das variáveis aleatórias, para todas as estruturas da matriz de correlações. Na medida em que o valor absoluto da correlação aumentava observou-se que a diferença absoluta entre as correlações desejadas e as correlações obtidas por meio dos valores simulados também aumentou. Observou-se ainda que as diferenças absolutas entre as correlações desejadas e as correlações obtidas foram maiores quando estas correlações eram negativas se comparadas quando estas eram positivas.

Afim de analisar a relação existente entre a correlação utilizada na distribuição normal multivariada (ρ) e a correlação obtida (ρ_{est}) entre os valores simulados das variáveis aleatórias que seguem distribuição Poisson e foram gerados pelo procedimento Norta, utilizou-se a função não-linear nos parâmetros, descrita abaixo:

$$\rho_{est} = \frac{\alpha}{1 + \beta e^{-\kappa\rho}} - \frac{\alpha}{1 + \beta} + e \tag{67}$$

em que α, β e κ são os parâmetros do modelo. Estes foram estimados utilizando o método de Gauss Newton e assim, obteve o seguinte modelo:

$$\hat{\rho_{est}} = \frac{3,7108}{1 + 1,3413e^{-0,8673\rho}} - 1,58493 \tag{68}$$

cujo gráfico é apresentado na Figura 15.

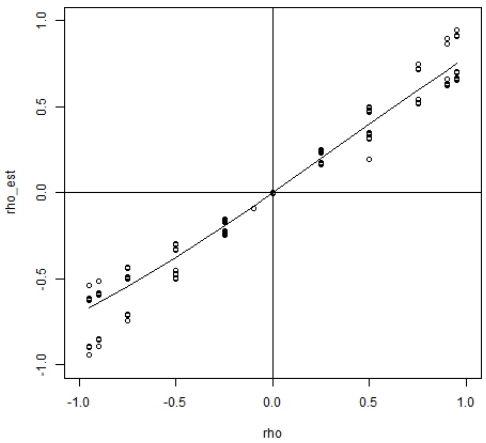

Figura 15 - Ajuste de um modelo não-linear nos parâmetros para descrever a relação entre a correlação dos vetores gerados pela simulação, denominado no gráfico de rho_est (ρ_{est}), e a correlação utilizada na simulação, denominada no gráfico de rho (ρ)

Com isso, a partir do modelo (68) é possível estimar as correlações que devem ser usadas na matriz de correlações da distribuição normal multivariada para obter as correlações desejadas entre os vetores aleatórios que seguem distribuição Poisson.

4.2 Exemplo de Aplicação

4.2.1 Análise da taxa de falsos alertas do aumento populacional de percevejos na *Glycine max L* (soja)

Os principais insetos-praga que atacam a cultura da *Glycine max L* (soja) na região do Distrito Federal são os percevejos *Euschistus heros* (Fabricius), *Nezara viridula* (Linnaeus) e *Piezodorus guildinii* (Fabricius) (Heteroptera: Pentatomidae) e a lagarta *Anticarsia gemmatalis* Hübner (Lepidoptera: Noctuidae). Dentre os percevejos, as espécies *E. heros* e P. *guildinii* têm apresentado, nos últimos anos, as mais altas densidades populacionais. Os percevejos surgem com o início da floração (fase reprodutiva), causando danos desde a formação das vagens até o final do desenvolvimento das sementes (maturidade fisiológica) (Fehr et al. 1971). O hábito de alimentação, sugando diretamente os grãos da soja, afeta o rendimento e a qualidade das sementes, exigindo que medidas de controle sejam adotadas quando as populações desses percevejos crescem a níveis muito elevados.

A partir dessas informações, utilizando modelagem matemática, é possível

estabelecer a relação entre a quantidade de indivíduos coletados nas armadilhas e a densidade absoluta de percevejos na área plantada, além de relacionar as coletas com o dano causado pelos percevejos na soja. Essa modelagem é uma ferramenta essencial que auxilia na tomada de decisão para o manejo da praga.

Considere um sistema que monitora a população de percevejos a partir do números de indivíduos capturados nas armadilhas em uma determinada plantação. Como o número de percevejos capturados é ilimitado, do ponto de vista teórico, assume-se que o número de percevejos dentro de cada armadilha segue uma distribuição de Poisson, com médias proporcional ao tamanho do local da plantação. Devido à proximidade geográfica, é razoável supor que os números de percevejos nas armadilhas em áreas geográficas próximas estão correlacionados. Dessa forma, tem-se vetores simulados que seguem distribuição Poisson n-variada, em que n representa o número de armadilhas na plantação e consequentemente a dimensão desse vetor. Para simplificar, supôs-se que o coeficiente de correlação do número de percevejos de cada par de armadilhas era igual a ρ e que todas as armadilhas capturavam em média o mesmo número λ de insetos.

Usou-se o algoritmo apresentado por Follmann (1996) e utilizado por Yahav e Shmueli (2012) para monitorar a série de capturas. Follmann (1996) apresenta uma técnica relativamente simples a partir da estatística multivariada de Mahalanobis no controle do aumento médio de uma ou mais séries. A estatística de monitoramento é dada por

$$\chi_t^2 = (\boldsymbol{X}_t - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{X}_t - \boldsymbol{\mu})$$

em que $\boldsymbol{X}_t$ é o vetor de contagem diária (de percevejos, no nosso exemplo) no tempo t; $\boldsymbol{\mu}$ é o vetor de média e $\boldsymbol{\Sigma}$ é a matriz de covariâncias. Um alerta é acionado quando

$$\left(\chi_t^2 > \chi_\alpha^2(n) \text{ e } \sum_{j=1}^{n} (X_t^j - \mu^j) > 0 \right).$$

Follmann (1996) prova que o procedimento tem taxa de falso alerta igual a α, e utiliza simulações para ilustrar a taxa de alerta verdadeira.

Simulou-se amostras do número de percevejos capturados variando o número n de armadilhas, o vetor de médias de insetos $\boldsymbol{\Lambda}$ e a correlação ρ, por um período de ausência de super população. Usou-se o método de Follmann (1996) com o nível de significância de $\alpha = 0,05$. Isso significa que permite-se 5% de falsos alertas (em média de 1 a 2 falsos alertas por mês). Sob essa configuração, o algoritmo deve idealmente produzir não mais

do que 5% de alertas (todos falsos, devido à falta de super população neste período). Os Gráficos 16, 17, 18 e 19 resumem a taxa de falsos alertas nos experimentos, quando os dados foram gerados da distribuição de Poisson multivariada.

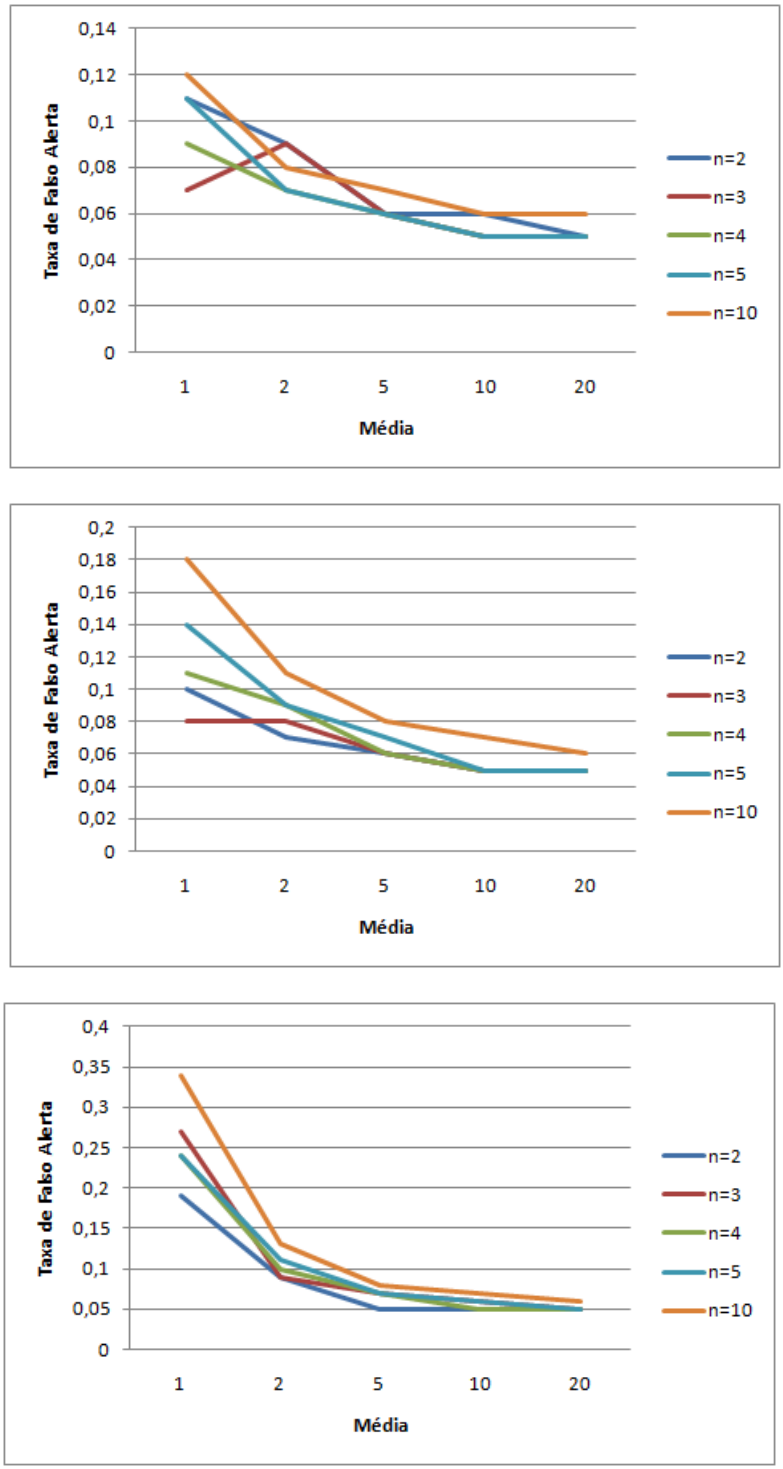

Figura 16 - Taxa de Falsos Alertas em relação a média quando $\rho = 0$, $\rho = 0,5$ e $\rho = 0,9$ e $n = 2, 3, 4, 5$ e 10

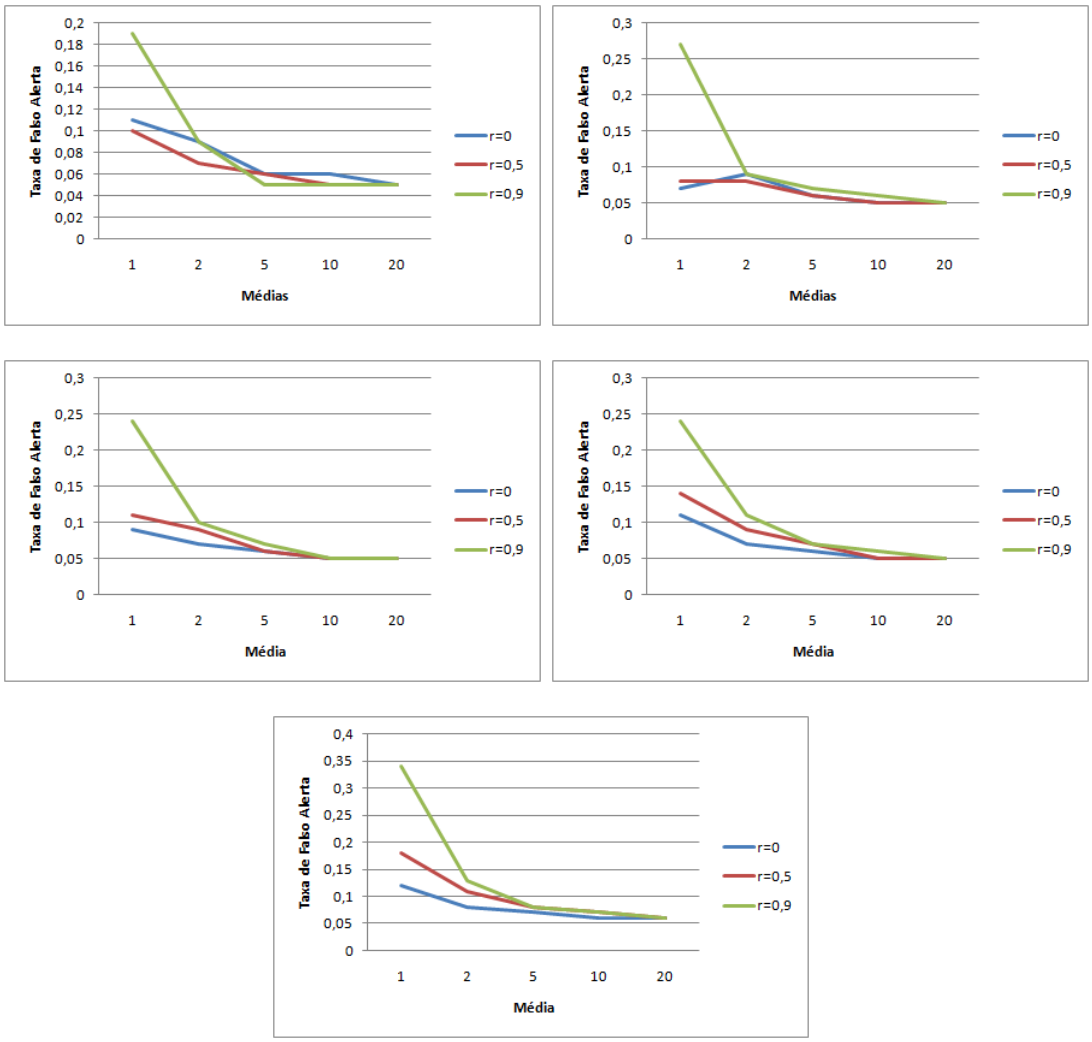

Figura 17 - Taxa de Falsos Alertas em relação a média para $n = 2$, $n = 3$, $n = 4$, $n = 5$, e $n = 10$ e $\rho = 0; 0,5$ e $0,9$

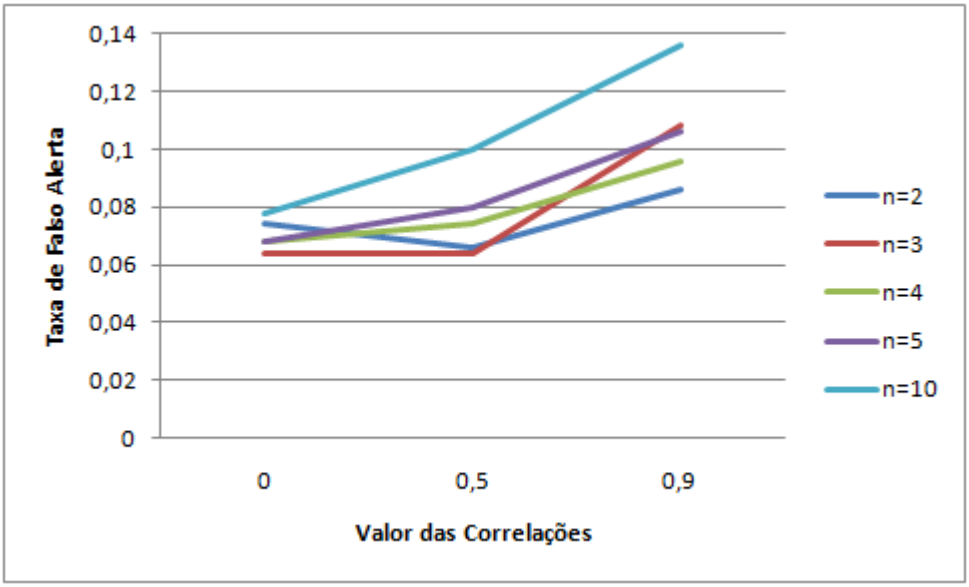

Figura 18 - Taxa de Falsos Alertas em função das correlações para número de armadilhas (n) diferentes

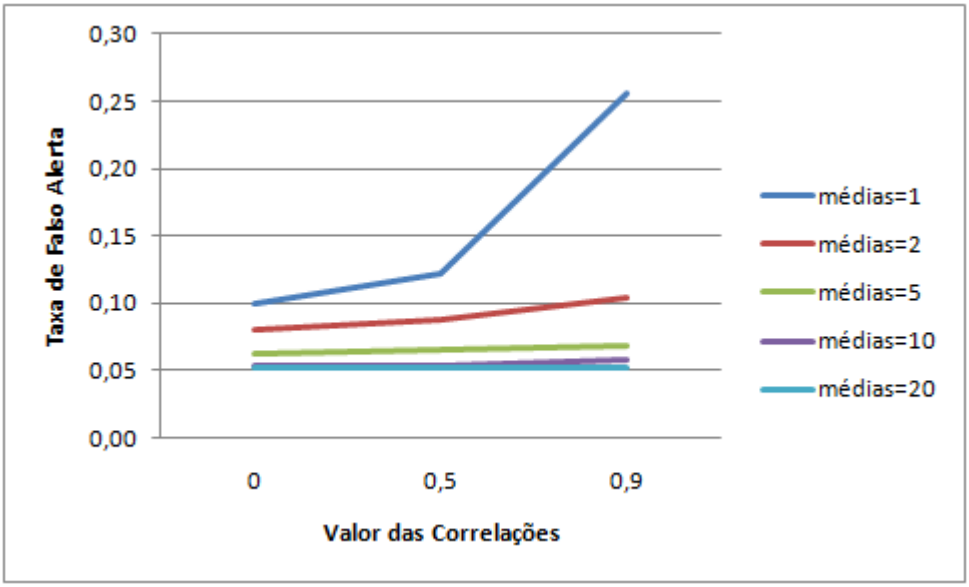

Figura 19 - Taxa de Falsos Alertas em função das correlações para médias diferentes

A partir das Figuras 16, 17, 18 e 19, observou-se que o método de Mahalanobis é sensível à distribuição utilizada. A taxa de falso alerta aumenta significativamente quando os dados seguem distribuição de Poisson, comparada com a taxa de quando são provenientes da distribuição normal. Em alguns casos, a taxa de falso alerta atinge 34%, ou seja, aproximadamente um em cada três alertas é falso. Na prática, altas taxas de

falsos alertas muitas vezes levam muitos usuários a ignorá-los, levando a um descrédito do sistema.

A taxa da falsos alertas aumentou quando a média λ diminuiu ou quando o número de armadilhas (n) aumentou (Figuras 16 e 18) e uma distribuição normal multivariada não pode mais ser usada como uma aproximação da distribuição de Poisson multivariada. Notou-se também que a taxa aumentou quando a correlação entre os números de percevejos capturados nas armadilhas aumentou (Figuras 17, 18 e 19). Diante disso, houve a necessidade de ajustar um modelo afim de estimar a taxa de falsos alertas em função das variáveis regressoras.

4.2.1.1 Regressão linear múltipla para a taxa de falsos alertas

Inicialmente, seguindo o modelo (40), utilizou-se o modelo de regressão linear múltipla para a variável resposta formada pela razão entre a taxa de falsos alertas supondo distribuição normal (α) e a obtida a partir da distribuição Poisson (α^*), juntamente com as variáveis regressoras λ, n e ρ, em que λ é o valor das médias, n é o número de armadilhas e ρ é o valor das correlações, cujo modelo foi:

$$\frac{\alpha}{\alpha^*} = \beta_0 + \beta_1 n + \beta_2 \rho + \beta_3 \lambda + \beta_4 n^2 + \beta_5 \rho^2 + \beta_6 \lambda^2 + e \tag{69}$$

4.2.1.2 Método de seleção de variáveis

Como as pressuposições necessárias para a realização da análise de regressão não foram atendidas realizou-se a transformação $\texttt{arcsen}\left(\sqrt{\frac{\alpha}{\alpha^*}}\right)$, em que $\texttt{arcsen}$ é a função arco seno. Dessa forma realizou-se análise de regressão para a variável transformada e utilizando os métodos de seleção de variáveis obteve-se, considerando o maior valor de $\overline{R}^2$, o valor da estatística de Mallows e o menor AIC pelo método do passo a frente passo atrás (stepwise), o seguinte resultado:

$$\texttt{arcsen}\left(\sqrt{\frac{\alpha}{\alpha^*}}\right) = \beta_0 + \beta_3 \lambda + \beta_4 n^2 + \beta_5 \rho^2 + \beta_6 \lambda^2 + e \tag{70}$$

4.2.1.3 Análise de Resíduos

Utilizando o modelo (70), verificou-se a qualidade do ajuste por meio da análise dos resíduos. Na Figura 20 apresenta-se o gráfico dos valores preditos e resíduos

estudentizados externamente.

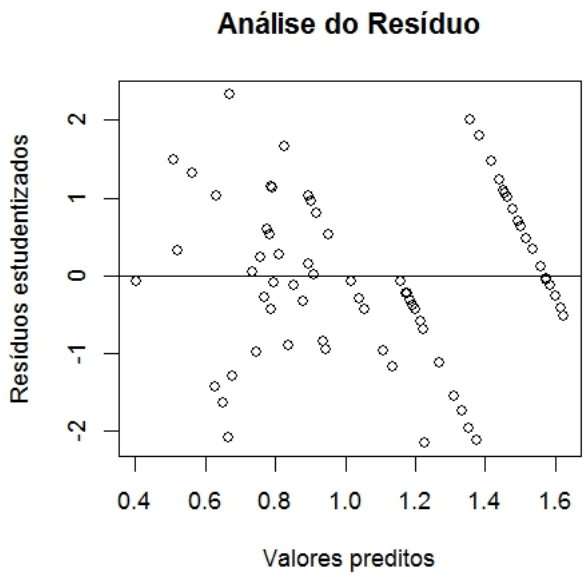

Figura 20 - Gráfico dos valores preditos e resíduos estudentizados externamente

Observa-se, na Figura 20, uma distribuição aleatória dos resíduos. Assim, visualmente, não há evidências de que os resíduos não sejam homogêneos. Na Figura 21 tem-se o gráfico normal de probabilidades (*Q-Q plot*) dos resíduos estudentizados externamente.

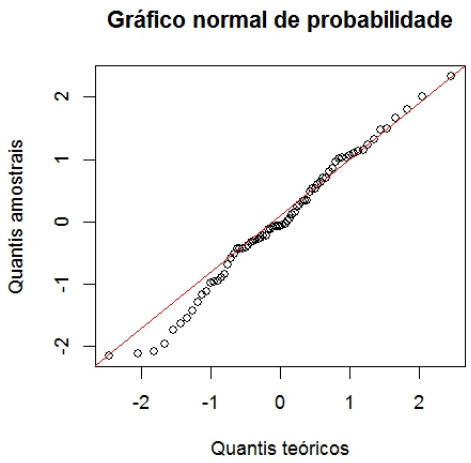

Figura 21 - Gráfico normal de probabilidades dos resíduos estudentizados do modelo de selecionado

Observa-se no gráfico normal de probabilidades uma dispersão nos extremos

e uma concentração dos dados na área central, área de maior ocorrência. A hipótese de normalidade dos resíduos, verificada pelo teste de *Shapiro-Wilk*, não foi rejeitada ($p-$valor $= 0,69$). Assim, não existem evidências para afirmar que a distribuição dos resíduos não seja normal. O histograma dos resíduos estudentizados externamente com a curva da distribuição normal estão na Figura 22.

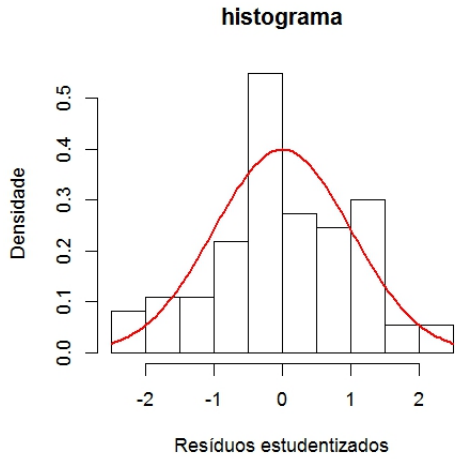

Figura 22 - Histograma dos resíduos estudentizados externamente e curva da distribuição normal

Na Figura 23 tem-se o gráfico de caixas (*box plot*) dos resíduos estudentizados externamente. Observa-se que não houve pontos discrepantes.

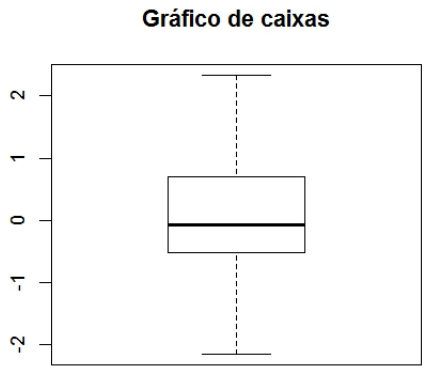

Figura 23 - Gráfico de caixas (*Box plot*) dos resíduos estudentizados externamente

4.2.1.4 Estimação dos parâmetros e análise de variância

Após a verificação das suposições do modelo de regressão (70), realizou-se o ajuste pelo método dos mínimos quadrados. As estimativas dos coeficientes de regressão e os respectivos erros padrão são apresentados na Tabela 2.

Tabela 2 - Estimativas dos coeficientes da regressão e os respectivos erros padrão

Parâmetro	Estimativa	Erro Padrão	Valor t	$Pr > \|t\|$
β_0	0,74218	0,03268	22,714	<0,001
β_3	0,11960	0,00804	14,867	<0,001
β_4	-0,00291	0,00037	-7,900	<0,001
β_5	-0,20033	0,03886	-5,154	<0,001
β_6	0,00372	0,00037	-10,064	<0,001

Dessa forma, a equação de regressão ajustada, do modelo (70) é dada por:

$$\widehat{\texttt{arcsen}\left(\sqrt{\frac{\alpha}{\alpha^*}}\right)} = 0,74218 + 0,11960\lambda - 0,00291n^2 - 0,20033\rho^2 - 0,00372\lambda^2$$

com coeficiente de determinação ajustado $\bar{R}^2 = 0,9072$.

Depois de ajustado o modelo, desfez-se a transformação obtendo o modelo

$$\widehat{\alpha^*} = \frac{\widehat{\alpha}}{\left(\texttt{sen}\ (0,74218 + 0,11960\lambda - 0,00291n^2 - 0,20033\rho^2 - 0,00372\lambda^2)\right)^2}$$

Dessa forma com base na correlação existente, no número de armadilhas, na média e na taxa de falsos alertas permitida, caso o vetor seguisse distribuição normal multivariada, é possível estimar a taxa de falsos alertas permitida, dado que o vetor segue distribuição Poisson Multivariada.

5 CONCLUSÕES

Observou-se que não existe problema no método de simulação NORTA, pois a distribuição Poisson se ajustou bem, para todos os vetores de médias e todas as estruturas da matriz de correlações, aos valores simulados das variáveis aleatórias, ao nível de 0,05 de significância. Concluindo-se com isso que as marginais dos dados trivariados simulados seguem distribuição Poisson. Mostrou-se que o método funciona bem e é altamente preciso na geração de dados multivariados com distribuições marginais de Poisson, para diferentes estruturas de correlações (negativas e positivas e variando os valores) e para médias λ assumindo valores altos e baixos.

Na medida em que o valor absoluto da correlação aumentava observou-se que a diferença absoluta entre as correlações desejadas e as correlações obtidas por meio dos valores simulados também aumentou, e que as diferenças absolutas entre as correlações desejadas e as correlações obtidas foram maiores quando estas correlações eram negativas se comparadas quando estas eram positivas. Conforme as médias aumentaram, a diferença entre as correlações desejadas e as correlações obtidas entre os vetores simulados diminuíram, o que era esperado uma vez que quando a média aumenta a distribuição Poisson se aproxima da distribuição normal.

Foram demonstradas vantagens práticas de dados que seguem distribuição Poisson multivariada sobre a normal multivariada na detecção de super populações de percevejos. Simulações inadequadas podem levar a excesso de falsos alertas. Uma vez que os dados seguem distribuição Poisson multivariada, a taxa de falsos alertas pode ser maior do que a imaginada. Essa taxa pode ser estimada por um modelo ajustado. A mesma técnica pode ser aplicada em diversos problemas em várias áreas do conhecimento.

REFERÊNCIAS

AVRAMIDIS, A. N.; CHANNOUF, N; L'ECUYER, P. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary margins and normal-copula dependence. **INFORMS Journal on Computing**, Hanover, v. 21, n. 1, p. 88-106, 2009.

BARR, D. R.; SLEZAK, N. L. A comparison of multivariate normal generators. **Communications of the Association for Computing Machinery**, New York, v. 15, n. 12, p. 1048-1049, 1972.

BARTLE, R. G. **The Elements of Real Analysis**. New York: John Wiley & Sons, 1976. 480p.

BATES, D. M.; WATTS, D. G. Relative curvature measures of nonlinearity (with discussion). **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 42, n. 1, p. 1-25, 1980.

BATES, D. M.; WATTS, D. G. **Nonlinear regression analysis and its applications**. New York: Wiley series in probability and mathematical statistics, 1988. 365p.

BEALE, E. M. L. Confidence regions in nonlinear estimation (with discussion). **Journal of the Royal Statistical Society. Series B (Methodological)**, London, v. 22, p. 41-88, 1960.

BOX, G. E. P.; LUCAS, H. L. Design of experiments in nonlinear situations. **Biometrika**, Oxford, n, 46, p. 77-90, 1959.

CARIO, M. C.; NELSON, B. L. **Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix.** Evanston: Northwestern University, Department of Industrial Engineering and Management Sciences, 1997. 19 p. Technical report.

CARIO, M. C.; B. L. NELSON. Numerical methods for fitting and simulating autoregressive-to-anything processes. **INFORMS Journal on Computing**, Hanover, v. 10, n. 1, p. 72-81, 1998.

CHEN, H. Initialização for NORTA: Generation of random vectors with specified marginals and correlations. **INFORMS Journal on Computing**, Hanover, v. 13, n. 4, p. 312-331, 2001.

CORDEIRO, G.M.; LIMA NETO, E.A. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, nº:16, 2004, Caxambu. **Modelos Paramétricos.** Caxambu: ABE, 2004. 246p.

DEMÉTRIO, C.G.B.; ZOCCHI, S.S. **Modelos de regressão**. Disponível em: <http://verde.esalq.usp.br/ jorge/cursos/clarice/ApostilaP.pdf> Acesso em: 20 ago. 2010.

DIAS, C. T. S.; SAMARANAYAKA, A.; MANLY, B. On the use correlated beta random variables with animal population modelling. **Ecological Modelling**, Amsterdã, v. 215, n.4, p. 293-300, 2008.

DOWNING,D. CLARK, J. **Estatística Aplicada**. Tradução de A. A. Farias, 2. ed. São Paulo: Saraiva, 2002. 351p.

DRAPER, N. R., SMITH, H. **Applied regression analysis**. 3.ed. New York: John Wiley & Sons Inc, 1998. 706p.

FARLIE, D. J. G. The performance of some correlations coefficients for a general bivariate distribution. **Biometrika**, Oxford, v. 47, p. 307-323, 1960.

FEHR, W. R; CAVINES, C. E.; BURMOOD, D. T.; PENNINGTON, J. S. Stage of development descriptions for soybeans, *Glycine max* (L.) Merrill. **Crop Science**, Madison, v. 11, p. 929-931, 1971.

FERREIRA, D. F. **Estatística Básica**. 2. ed rev. Lavras: UFLA, 2009. 664p.

FOLLMANN, D. A simple multivariate test for one-sided alternatives. **Journal of the American Statistical Association**, Alexandria, v. 91, n.434, p. 854-861, 1996.

FREUND, J. E.; SIMON, G. A. **Estatística Aplicada:** economia, administração e contabilidade. Tadução de A. A. Farias, 9. ed. Porto Alegre: Bookman, 2000. 404p.

GALLANT, A. R. **Nonlinear statistical models**. New York: Wiley series in probability e mathematical statistics, 1987. 624p.

GENEST, C. Frank's family of bivariate distributions. **Biometrika**, Oxford, v. 74, p. 549-555, 1987.

GENEST, C.; MACKAY, R. J. The joy of copulas: Bivariate distributions with uniform marginals. **The American Statistician**, Washington, v. 40, p. 280-283, 1986.

GHOSH, S.; HENDERSON, S. G. Behavior of the NORTA method for correlated random vector generation as the dimension increases. **ACM Transactions on Modeling and Computer Simulation**, New York, v. 13, n. 3, p. 276–294, 2003.

GROSS, K.; LOCKWOOD, J. R.; FROST, G. G; MORRIS, W. F. Modeling controlled burning and trampling reduction for conservation of Hudsonia Montana. **Conservation Biology**, Sunderland, v. 12, p. 1291-1301, 1998.

HAHN, G. J.; SHAPIRO, S. S. **Statistical models in engineering**, New York:John Wiley & Sons, 1967. 355p

HAIR JR., J. F.; ANDERSON, R. E.; TATHAM, R. L. BLACK, W. C. **Análise Multivariada de Dados**. Tadução de A.S. Sant'anna e A. Cloves Neto, 5. ed. Porto Alegre: Brokman, 2006. 593p.

HERNDVLGYI, I. T. Generating random vectors from the multivariate normal distribution. Working paper. 1998. Disponível em: <http://citeseer.ist.psu.edu/78434.html>. Acesso em: 8 nov. 2010.

HURST, R. L.; KNOP, R. E. Algorithm 425. Generation of random correlated normal variables. **Communications of the Association for Computing Machinery**, New York, v. 15, p. 355-357, 1972.

JOHNSON, M. E. **Multivariate statistical simulation**, New York:John Wiley & Sons, 1987. 230p.

JOHNSON, M. E.; TENENBEIN, A. A bivariate distribution family with specified marginals. **Journal of the American Statistics Association**, Alexandria, v. 76, p.

198-201, 1981.

KARLIS, D. An EM algorithm for multivariate poisson distribution and related models. **Journal of Applied Statistics**, Oxon, v. 30, n. 1, p. 63-77, 2003.

KOEHLER, K. J.; SYMANOWSKI, J. T. Constructing multivariate distributions with specific marginal distributions. **Journal of Multivariate Analysis**, New York, v.55, p. 261-282, 1995.

KRUMMENAUER, F. Limit theorems for multivariate discrete distributions. **Metrika**, London, v. 47, n. 1, p. 47-69, 1998.

KRUMMENAUER, F. Efficient simulation of multivariate binomial and poisson distributions. **Biometrical Journal**, Weinheim, v. 40, n. 7, p. 823-832, 1999

MALLOWS, C. L. Some Comments on C_p. **Technometrics**, Milwaukee, v. 15, n.4, p. 661-675, 1973

MARDIA, K. V. **Families of Bivariate Distributions**, London: Griffin, 1970. 109p.

MARSHALL, A. W.; OLKIN, I. Families of multivariate distributions. **Journal of the American Statistics Association**, Alexandria, v. 83, n. 403, p. 834-841, 1988.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada:** uma abordagem aplicada. Belo Horizonte: UFMG, 2005. 295p

MINHAJUDDIN, A. T.; HARRIS, I. R.; SCHUCANY, W. Simulating multivariate distributions with specific correlations.**Journal of Statistical Computation and Simulation**, New York, v. 74, n. 81, p 599-607, 2004.

MONTGOMERY, D.C.; PECK, E.A.; VINING, G.G. **Introdution to linear regression analysis**. 4.ed. New York: John Wiley & Sons, 2006. 612p.

MONTGOMERY, D.C.; PECK, E.A. **Introdution to linear regression analysis**. 2.ed. New York: John Wiley & Sons, 1992. 527p.

NASH, J. C. **Compact Numerical Methods for Computers:** Linear Algebra and Function Minimisation. Bristol: Adam Hilger, 1979. 281p.

PARK, C. P. Construction of random vectors of heterogeneous component variables under specified correlation structures. **Computational Statistics and Data Analysis**, Amsterdã, v. 46, p. 621-630, 2004.

PLACKETT, R. L. A class of bivariate distributions. **Journal of the American Statistics Association**, Alexandria, v. 60, p. 516-522, 1965.

R DEVELOPMENT CORE TEAM. **R:** A language and environment for statistical computing. **R Foundation for Statistical Computing**, Austria, 2010.

RATKOWSKY, D.A. **Nonlinear regression modelling**. New York: Marcel Dekker, 1983. 241p.

RIPLEY, B. D. **Stochastic Simulation**, New York:John Wiley & Sons, 1987. 237p.

ROYDEN, H. L. **Real Analysis**, New York: Prentice Hall, 3 ed., 1988. 444p.

SARTORIO, S. D.**Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R.** 2008. 130p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2008.

SCHEUR, E. M.; STOLLER, D. S. On the generation of normal random vectors. **Technometrics**, Alexandria, v. 4, n. 2, p. 278-281, 1962.

SHIN, K.; PASUPATHY, R. An Method for Fast Generation of Bivariate Poisson Random Vectors. **Proceeding of the 2007 Winter Simulation Conference**, 2007. p. 472-479.

SHIN, K.; PASUPATHY, R. An Algorithm for Fast Generation of Bivariate Poisson Random Vectors. **INFORMS Journal on Computing**, Hanover, v. 22, n. 1, p. 81-92, 2010.

WHITT, W. Bivariate distributions with given marginals. **The annals of Statistics**, Philadelphia, v. 4, n. 6, p. 1280–1289, 1976.

WILKS, S. S. **Mathematical Statistics**, New York:John Wiley & Sons, 1962. 644p.

YAHAV, I.; SHMUELI, G. **An Elegant Method for Generating Multivariate Poisson Random Variables**. arXiv:0710.5670v2 [stat.CO], 2008.

YAHAV, I.; SHMUELI, G. On generating multivariate Poisson data in management science applications. **Applied Stochastic Models in Business and Industry**, New York, v. 28, n. 1, p. 91-102, 2012.

YUE, S.; QUARDA, T. B. M. J.; BOBÉE, B. A review of bivariate gamma distributions for hydrological applications. **Journal of Hydrology**, Philadelphia, v. 246, p 1-18, 2001.

ANEXOS

ANEXO A - Programa (R) utilizado na simulação de variáveis aleatórias com distribuição Poisson trivariada a partir de uma distribuição normal trivariada com vetor de médias μ e matriz de dispersão (Σ) (procedimento Norta).

```
sink("C:/Users/Doutorado/Simulação/lam5_corr_dif1.txt",
     append=TRUE, split=TRUE)

#Limpa a memória do R #
rm(list=ls(all=TRUE))

# Ativa a biblioteca abaixo, deixando-a pronta para uso $
library(MASS)

# Registra o tempo de simulação $
ptm <- proc.time()



## Define os parâmetros e as matrizes usadas nas simulações ##

rho11=1; rho22=1; rho33=1; rho12=0.5; rho13=-0.5; rho23=0.25
rho <- cbind(c(rho11, rho12, rho13), c(rho12, rho22,rho23),
             c(rho13,rho23,rho33))
(eigen(rho))$values
mi<-c(0,0,0); lambda=c(5,5,5)
n_sim=1000
n_elem=500
n_variaveis=3
Z1=matrix(,n_elem,1); z=matrix(,n_elem,1);P=matrix(,n_elem,n_variaveis)
corr_P1P2<-c(); corr_P1P3<-c(); corr_P2P3<-c()
cov_P1P2<-c(); cov_P1P3<-c() ; cov_P2P3<-c()
media_pois<-matrix(,n_sim,n_variaveis);var_pois<-matrix(,n_sim,n_variaveis)
```

```
## Gera uma amostra aleatória com distribuição Poisson trivariada
#  a partir da distribuição normal trivariada ##

# Calcula as médias das distribuições marginais e as correlações entre elas #

for (l in 1:n_sim) {
x <- mvrnorm(n_elem, mu=mi, Sigma=rho)
U1<-pnorm(x[,1], mean = mi[1], sd = 1)
U2<-pnorm(x[,2], mean = mi[2], sd = 1)
U3<-pnorm(x[,3], mean = mi[3], sd = 1)
u<-matrix(c(U1,U2,U3),nrow=n_elem,ncol=n_variaveis, byrow=FALSE)

for (j in 1:n_variaveis) {
for (i in 1:n_elem) {
for (k in 1:n_elem) {
Z1[k]=dpois(k-1,lambda[j])
z[k]=sum(Z1[1:k])
if(   any(z[k] < u[i,j])) P[i,j] <- k else stop
}}}
corr_P1P2[l]<- cor(P[,1],P[,2]); corr_P1P3[l]<- cor(P[,1],P[,3]);
corr_P2P3[l]<- cor(P[,2],P[,3])
cov_P1P2[l]<- cov(P[,1],P[,2]); cov_P1P3[l]<- cov(P[,1],P[,3]);
cov_P2P3[l]<- cov(P[,2],P[,3])
media_pois[l,]<- t(c(mean(P[,1]),mean(P[,2]), mean(P[,3])))
var_pois[l,]<- t(c(var(P[,1]),var(P[,2]),var(P[,3]))) }


corr_P1P2; mean(corr_P1P2)
corr_P1P3; mean(corr_P1P3)
corr_P2P3; mean(corr_P2P3])
cov_P1P2; mean(cov_P1P2)
cov_P1P3; mean(cov_P1P3)
```

```
cov_P2P3; mean(cov_P2P3)
media_pois; mean(media_pois)
var_pois; mean(var_pois)
(correlacoes=c(mean(corr_P1P2),mean(corr_P1P3),mean(corr_P2P3)))
proc.time()- ptm
sink()
write.csv2(P, file="C:/Users/Doutorado/Simulação/lam5_corr_dif1.csv")
```

ANEXO B - Programa (R) utilizado para testar se a distribuição Poisson se ajustava aos vetores aleatórios gerados pelo procedimento Norta e criar o gráfico de frequência observada *vs* frequência esperada.

```
## Teste de aderência qui-quadrado ##

# Limpa a memória do R#

rm(list=ls(all=TRUE))

Entra com os dados gerados e salvos do endereço abaixo#

dados=read.csv2("C:/Users/Doutorado/Simulação/lamb5_corr_dif1.csv", header=TRUE)
dados

#Calcula a frequência observada#

dad=dados$V1
(f<-c(0,table(dad)))
(n=length(f))
(x_i=0:(n-1))
names(f) <- c(x_i[-n],">12")
f

#Calcula a frequência esperada#

(pi <- dpois(x_i[-n], lambda=5))
(pi <- c(pi, 1-sum(pi)) )
(f_ast <- sum(f)*pi)

#Teste de aderência qui-quadrado#

(X2 <- sum((f-f_ast)^2/f_ast))
```

```
(nu <- length(f)-1 -1 )
pchisq(X2, df=nu, lower.tail=FALSE)

#Salva o gráfico de frequência observada versus frequência esperada#

jpeg(filename = "lamb5cor_dif1_P1.png", width = 480, height = 480,
     units = "px", pointsize = 12, quality = 100,
     bg = "white",  res = NA, restoreConsole = TRUE)
plot(x_i,f,type="h",xlab="P", ylab="frequência")
lines(x_i,f_ast,col = "red",lwd=2)
legend("topright", legend=c("frequência observada",
"frequência esperada"), lty=c(1,1), col=c(1,2), lwd=1:2, bty="n", pch=c(NA,NA))
dev.off()
```

ANEXO C - Programa (R) para obter as estimativas dos parâmetros do modelo não linear nos parâmetros, utilizando o método de Gauss-Newton

```
# Limpa a memória do R#

rm(list=ls(all=TRUE))

Entra com os dados gerados e salvos do endereço abaixo#

dados=read.csv2("C:/Users/Doutorado/Simulação/dad_estimacao.csv", header=TRUE)
dados
attach(dados)
names(dados)

# Encontra os valores iniciais para o método numérico

plot(rho,rho_est,xlim = c(-1,1), ylim=c(-1,1))
abline(h=0,v=0)
x=seq(-1,1,0.01)
alpha=14.34
beta=4.86
k=0.32
sig=alpha/(1+beta*exp(-k*x))-alpha/(1+beta)
lines(x,sig)
y=dados$rho_est
x=dados$rho
cbind(x,y)

# Ajusta o modelo (utiliza a função nls do R)

funcao=y~theta1/(1 + theta2*exp(-theta3*x))-theta1/(1+theta2);
func=nls(funcao, start = list(theta1 = 14.345, theta2 = 4.87, theta3 = 0.32) )
```

```
resul=summary(func)
resul
attach(resul)

# Escreve o modelo ajustado e faz o gráfico #

y=parameters[1,1]/(1 + parameters[2,1]*exp(-parameters[3,1]*x))-
(parameters[1,1]/(1+resul$parameters[2,1])) ;
cbind(x,y)
plot(rho,rho_est,xlim = c(-1,1), ylim=c(-1,1))
abline(h=0,v=0)
lines(x,y)
```

ANEXO D - Programa (R) utilizado na análise de regressão linear múltipla para a taxa de falsos alertas

```
# Entrada de dados #
dados=read.csv2("C:/Users/Raphael/Desktop/resultado exemplo soja.csv")

# retira o ponto discrepante #
dados=dados[c(-4,-13),]

# Renomeia e define as variáveis independentes do modelo #
dados
names(dados)
dados$X1=dados$n
dados$X2=dados$rho
dados$X3=dados$lambda
dados$X4=dados$X1^2
dados$X5=dados$X2^2
dados$X6=dados$X3^2

# Define a variável Dependente
dados$Y=dados$rel_alphas

# Define a variável dependente transformada
dados$Y=asin(sqrt(dados$rel_alphas))
attach(dados)

## Definição do modelo e Análise de Regressão ##
# Define o modelo completo #

mod <- lm(Y~X1+X2+X3+X4+X5+X6)
mod

# Análise de Regressão Linear Múltipla e obtenção do R2 #
```

```
anova(mod)
summary.lm(mod)

## Diagnósticos ##

# Gera os resíduos estudentizados #
rs <- rstudent(mod)

# Gera os gráficos de diagnósticos #
hist(rs, freq = FALSE, main="histograma", xlab = "Resíduos estudentizados",
ylab= "Densidade")
curve(dnorm(x,0,1), col = 2, lty = 1, lwd = 2, add = TRUE)

boxplot(rs, main="boxplot")

qqnorm(rs, xlab = "Quantis teóricos", ylab = "Quantis amostrais")
qqline(rs,col="red")

require(MASS)
boxcox(mod,lambda = seq(0, 3, 1/10))

# Gráfico para Análise de Resíduos #

# Gráfico dos valores preditos versus os resíduos estudentizados #

yp <- predict.lm(mod)
par(mfrow=c(1,1))
plot(yp,rs,main="Análise do Resíduo", xlab="Valores preditos",
ylab="Resíduos estudentizados")
abline(h=0)

# Gráfico para val. influentestes - Laverage (hat)#
```

```
N=length(Y); P=length(mod$coefficient)
h <- lm.influence(mod)$hat; lc <- 3*P/N
minrs=min(min(rs),-3); maxrs=max(max(rs),3); ymin=minrs-.1; ymax=maxrs+.1
maxh=max(max(h),lc); minh=min(h); xmin=minh-.1; xmax=maxh+.1
par(mfrow=c(1,1))
plot(c(xmin,xmax),c(ymin,ymax), type="n", xlab="h - leverage",
ylab="Resíduos estudentizados")
abline(h=-3, col="red"); abline(h=3,col="red");
abline(v=lc, col="blue"); points(h,rs)


# faz o testes de normalidade de Shapiro Wilk #

shapiro.test(rs)

## Se for o caso, elimine as observações (os registros) e reinicie a análise ##

## Seleção de Modelos ##

#  Método Cp (Mallows´Cp) - Menor Cp - melhor o modelo) #
# Método R2 Ajustado - Maior R2 - melhor o modelo) #

install.packages("leaps")
require(leaps)

# Modelo completo com todas as variáveis explanatórias #

selr2 <- leaps(x=cbind(X1,X2,X3,X4,X5,X6), y=Y,method=c("Cp"))
nparr2 <- selr2$size
r2.aj <- selr2$adjr2
dfr <- data.frame(cbind(nparr2,r2.aj,selr2$which))
sdfr <- dfr[order(dfr$r2.aj,decreasing="T"),]
```

```
sdfr

# Método backward ou forward ou stepwise $
#a seleção é feita pelo valor de AIC (Critério Akaike)

step(mod, direction="backward")   # método (backward).
step(mod, direction="forward")   # método (forward).
step(mod, direction="both")       # método (stepwise).

## Análise com o Modelo selecionado ##

mod <- lm(Y~X3+X4+X5+X6)
summary(mod)

#teste do modelo ajustado
#alpha: velor desejado para a TFA (taxa de falso alerta)

yp <- predict.lm(mod)
ypredito=mod$coef[1]+mod$coef[2]*X3+mod$coef[3]*X4+mod$coef[4]*X5
+mod$coef[5]*X6
yp=ypredito
yp=(sin(yp))^2
TFA_est=0.05/yp
TFA_est

# Diagrama de dispersão com o gráfico do modelo ajustado #
plot(Poisson,TFA_est)
lines(seq(0,0.35,0.01),seq(0,0.35,0.01))
```