**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Regiões de confiança para a localização do ponto estacionário em
superfícies de resposta, usando o método “bootstrap” Bayesiano**

David José Miquelluti

Tese apresentada para obtenção do título de Doutor em
Agronomia. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba

2008

David José Miquelluti

Engenheiro Agrônomo

**Regiões de confiança para a localização do ponto estacionário em superfícies de
resposta, usando o método “bootstrap” Bayesiano**

Orientador:

Prof. Dr. **SILVIO SANDOVAL ZOCCHI**

Tese apresentada para obtenção do título de Doutor em
Agronomia. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba

2008

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Miquelluti, David José

Regiões de confiança para a localização do ponto estacionário em superfícies de reposta, usando o método "bootstrap" Bayesiano / David José Miquelluti.- - Piracicaba, 2008.

57 p.

Tese (Doutorado) - - Escola Superior de Agricultura Luiz de Queiroz, 2008.
Bibliografia.

1. Amendoim – Estatísticas e dados numéricos 2. Distribuições amostrais
3. Inferência bayesiana 4. Metodologia e técnicas de computação 5. Superfícies de resposta I.Título

CDD 519.54

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

Dedicatória

A DEUS,
fonte inspiradora e sede da eterna sabedoria que
tem dirigido minhas ações ao longo da vida.

Aos meus PAIS,
exemplo de luta e amor,
fundamentais para a minha formação.

A JUCLÉIA e ao DANIEL,
pelo amor, apoio incondicional e à imensa
compreensão em todos os momentos dessa etapa.

AGRADECIMENTOS

Ao meu orientador, Prof. Silvio Sandoval Zocchi, a orientação, o apoio e a amizade.

Aos professores do Departamento de Ciências Exatas da ESALQ/USP, os ensinamentos e a amizade.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, a atenção e a amizade.

Aos amigos e colegas de turma Ana Maria, Beth, Denise, Fogo, Geneville, Idemauro, Juliana, Luciano, Pedro, Alfredo, Francine, Hélio, Júlio, Melissa e Willian, os bons momentos, a amizade e a troca de conhecimentos.

Ao Cesar e ao Osmar, o acolhimento em minhas visitas a Piracicaba.

A todos os amigos do mestrado e doutorado, a convivência e a amizade.

A Universidade do Estado de Santa Catarina (UDESC), a liberação para o curso e o apoio financeiro.

Aos professores dos cursos de Agronomia e Medicina Veterinária da UDESC, o incentivo, a amizade e a confiança.

A todos, que de forma direta ou indireta, contribuíram para a realização deste trabalho.

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	9
1 INTRODUÇÃO	10
2 DESENVOLVIMENTO	12
2.1 Revisão de Literatura	12
2.1.1 Métodos bootstrap	12
2.1.2 Metodologia de superfícies de resposta	13
2.1.3 Regiões de confiança das coordenadas do ponto estacionário	17
2.2 Metodologia	19
2.2.1 Material	19
2.2.2 Métodos	19
2.2.2.1 Regiões de confiança	19
2.2.2.2 Simulação	30
2.2.2.3 Análise dos dados de produção de amendoim	32
2.3 Resultados	33
2.3.1 Simulação	33
2.3.2 Análise dos dados de produção de amendoim	37
3 CONSIDERAÇÕES FINAIS	41
REFERÊNCIAS	43
APÊNDICES	46

RESUMO

Regiões de confiança para a localização do ponto estacionário em superfícies de resposta, usando o método “bootstrap” Bayesiano

Experimentos nos quais uma ou mais variáveis respostas são influenciadas por diversos fatores quantitativos são bastante comuns nas áreas agrícola, química, biológica, dentre outras. Nesse caso, o problema de pesquisa consiste em se estudar essa relação, sendo de grande utilidade o uso da metodologia de superfícies de resposta (MSR). Nesse contexto, a determinação dos níveis dos fatores que otimizam a resposta consiste inicialmente na obtenção das coordenadas do ponto estacionário do modelo ajustado. No entanto, como o modelo verdadeiro é desconhecido, é interessante obter uma região de confiança das coordenadas verdadeiras de modo a avaliar a precisão da estimativa obtida. Foram abordados aqui os procedimentos para construção de regiões de confiança para as coordenadas do ponto estacionário em diferentes situações considerando-se a forma das superfícies analisadas e a distribuição e magnitude da variância dos erros do modelo. Foram utilizadas a metodologia de Box e Hunter (1954) (BH), “bootstrap” e “bootstrap” Bayesiano aliados ao cálculo da distância de Mahalanobis entre as coordenadas do ponto estacionário da amostra observada e aquelas obtidas por meio das estimativas “bootstrap” (BM e BBM), e métodos “bootstrap” e “bootstrap” Bayesiano aliados a métodos não paramétricos de estimação de funções densidade de probabilidade (BNP e BBNP). A avaliação da metodologia foi realizada por meio de simulação e foi aplicada a um conjunto de dados de produção de amendoim. No estudo de simulação, a metodologia BH, baseada na distribuição normal dos erros, apresentou um bom desempenho em todas as situações analisadas, havendo concordância entre as regiões de confiança nominais e reais, mesmo naquelas em que essa distribuição é bastante assimétrica. Este mesmo comportamento foi observado para os métodos BM e BBM. No entanto, os métodos BNP e BBNP não apresentaram um desempenho satisfatório, resultando em um nível de significância real menor que o nominal para os autovalores com menor valor absoluto, gerando regiões de confiança maiores. No caso de autovalores com maior valor absoluto observou-se situação inversa. No caso da análise do conjunto de dados de amendoim os métodos BH, BM e BNP apresentaram regiões de confiança mais amplas comparativamente aos métodos BBM e BBNP. No entanto, os valores das estimativas do “bootstrap” Bayesiano são mais próximas das estimativas de mínimos quadrados e apresentam menor dispersão o que explica a menor área da região de confiança.

Palavras-chave: Superfície de resposta; Ponto estacionário; Regiões de confiança; “Bootstrap” Bayesiano; Densidade não paramétrica

ABSTRACT

Confidence region on the location of the stationary point in response surfaces, a Bayesian bootstrap approach

Experiments in which one or more response variables are influenced by several quantitative factors are very common in agricultural, chemistry, biology and other areas. In this case, the research question consists in studying this relation, being of great utility the use of response surface methodology (RSM). In this context determining the level of the factors that optimize the response consists in finding the coordinates of the stationary point of the model. However, as the true model is unknown, it is of interest to obtain a confidence region of the true coordinates to analyze the precision of the obtained estimate. The procedures for the construction of confidence regions for the coordinates of the stationary point were studied in different situations, considering the shape of the surfaces analyzed and the distribution and magnitude of the variance errors. The methodology of Box and Hunter (1954) (BH), bootstrap and Bayesian bootstrap with Mahalanobis distance among the coordinates of the stationary point of the observed sample and those obtained using bootstrap estimates (BM and BBM) and bootstrap and Bayesian bootstrap with non-parametric methods for density estimation (BNP and BBNP) were compared. The methodology evaluation was realized by means of simulation and applied to a peanut yields data set. In simulation study the BH methodology, which is based in normal distribution of errors, presented a good performance in all of the analyzed situations, having concordance among the nominal and real confidence regions, even in those which this distribution is fairly asymmetric. This behavior was also observed in BM and BBM methods. The BNP and BBNP methods did not presented a satisfactory performance, resulting in a real significance level lower than the nominal for the eigenvalue with lower absolute value, generating bigger confidence regions. The inverse was observed using eigenvalue with higher absolute value. In the analysis of the peanut yields data set the BH, BM and BNP methods presented confidence regions larger than the BBM and BBNP methods. The Bayesian bootstrap estimate values are closer of the minimum square estimates and present less dispersion what explain the confidence region lower area.

Keywords: Respose surface; stationary point; Confidence regions; Bayesian bootstrap; Non parametric density

LISTA DE FIGURAS

Figura 1 - Regiões de 90% e de 95% de confiança para a localização do ponto esta-
cionário utilizando o método de Box e Hunter (1954) 38

Figura 2 - Regiões de 90% e de 95% de confiança para a localização do ponto esta-
cionário utilizando o método “bootstrap” com contornos da densidade não
paramétrica bivariada - $\mathbf{h} = [0,0311\ 0,0750]^T$ (esquerda) e da distância de
Mahalanobis (direita) considerando-se $M = 1000$ amostras 39

Figura 3 - Regiões de 90% e de 95% de confiança para a localização do ponto esta-
cionário utilizando o método “bootstrap” Bayesiano com contornos da den-
sidade não paramétrica bivariada - $\mathbf{h} = [0,0199\ 0,0310]^T$ (esquerda) e da
distância de Mahalanobis (direita) considerando-se $M = 1000$ amostras . . . 39

LISTA DE TABELAS

Tabela 1 - Produtividade de amendoim (libra/parcela) em função das quantidades (libra/parcela) de dois fertilizantes, NPK e SN. (rep.=repetição)	20
Tabela 2 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de B e com a distribuição dos erros - Metodologia de Box e Hunter (1954)	34
Tabela 3 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de B e com a distribuição dos erros - Distância de Mahalanobis	35
Tabela 4 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de B e com a distribuição dos erros - contorno de densidade não paramétrica bivariada . . .	36
Tabela 5 - Estimativas de θ e erros padrões obtidos através dos métodos de mínimos quadrados, “bootstrap” clássico e “bootstrap” Bayesiano	38

1 INTRODUÇÃO

Experimentos nos quais uma ou mais variáveis respostas são influenciadas por diversos fatores quantitativos são bastante comuns nas áreas agrícola, química, biológica, dentre outras. Nesse caso, o problema de pesquisa consiste em se estudar essa relação, sendo de grande utilidade o uso da metodologia de superfícies de resposta (MSR).

Apesar de as idéias básicas da MSR terem aparecido no século XIX, o grande impulso na análise de dados experimentais por meio de superfícies de resposta foi dado por Box e Wilson (1951) e nos artigos posteriores publicados por Box e seus associados.

A MSR compreende um conjunto de técnicas estatísticas que incluem desde o planejamento do experimento até a análise dos dados, com o objetivo de relacionar a(s) resposta(s) com os níveis de fatores quantitativos que a(s) afeta(m). Por exemplo, em agricultura, a resposta pode ser a produtividade de uma cultura e os fatores, as quantidades de diferentes tipos de nutrientes aplicados na lavoura.

Especificamente, a MSR tem como objetivos: otimizar a(s) resposta(s); encontrar e explorar a região do espaço de fatores que resulta em condições desejáveis do processo ou produto; conhecer a forma geral da superfície estudada com vistas a atingir os objetivos anteriores.

A determinação dos níveis dos fatores que otimizam a resposta consiste inicialmente na obtenção das coordenadas do ponto estacionário de um modelo ajustado, geralmente um modelo de segunda ordem completo. No entanto, devido à variação amostral inerente ao processo experimental, é interessante obter uma região de confiança para as coordenadas do ponto estacionário do modelo real desconhecido de modo a avaliar a precisão de suas estimativas. Nesse contexto, várias alternativas têm sido sugeridas para atingir tal objetivo, sendo a primeira delas, a metodologia apresentada por Box e Hunter (1954) que é baseada na pressuposição de normalidade dos erros aleatórios do modelo. Recentemente, Freitas (1993), Zocchi e Silva (2003) e Silva e Zocchi (2004) apresentaram formas diferentes de se construir as regiões de confiança baseadas na metodologia “bootstrap” não paramétrica.

“Bootstrap” é uma técnica computacionalmente intensiva utilizada para se estimar a distribuição amostral de uma estatística (MANLY, 1997). Dada uma amostra aleatória

de tamanho n , essa técnica consiste na reamostragem, com reposição, de uma função de distribuição não paramétrica que atribui probabilidades iguais a $\frac{1}{n}$ a todas as n observações. Como alternativa, Rubin (1981) apresenta uma variação dessa técnica por meio da modificação na forma de atribuição dessas probabilidades, utilizando para isso uma distribuição de probabilidades a priori. Segundo o autor, esse método seria o análogo Bayesiano do “bootstrap” clássico. Dadas as características deste, acredita-se que ele propicie uma maior cobertura do espaço paramétrico e com isto uma determinação mais precisa das regiões de confiança do parâmetro.

Neste trabalho, é proposta uma nova metodologia para determinação de regiões de confiança para as coordenadas do ponto estacionário de uma superfície de resposta representada por um modelo polinomial de segundo grau, baseada no conceito de “bootstrap” Bayesiano apresentado em Rubin (1981) aliada à técnica de estimação de densidades não paramétrica (SILVERMAN, 1986; BOWMAN e AZZALINI, 1997), descrita em Zocchi e Silva (2003) e Silva e Zocchi (2004). Os resultados são comparados com os obtidos por meio da utilização da metodologia de Box e Hunter (1954) e do método “bootstrap” seguido pelo cálculo da distância de Mahalanobis, apresentado por Freitas (1993). Os comportamentos dos diferentes métodos de obtenção das regiões de confiança são em seguida avaliados por meio de estudos de simulação. Um exemplo de aplicação ilustra a metodologia cujos procedimentos computacionais desenvolvidos foram implementados no programa R (R DEVELOPMENT CORE TEAM, 2006).

2 DESENVOLVIMENTO

2.1 Revisão de Literatura

Nesta seção é apresentada uma breve descrição sobre os métodos “bootstrap” clássico e Bayesiano e sobre a metodologia de superfícies de resposta, incluindo a obtenção de regiões de confiança das coordenadas do ponto estacionário de uma superfície.

2.1.1 Métodos bootstrap

Uma das tarefas da pesquisa quantitativa consiste em fazer inferências probabilísticas, sobre uma característica populacional (parâmetro), baseadas num estimador obtido a partir de uma amostra extraída da população de interesse. A forma paramétrica tradicional de inferência utiliza fórmulas analíticas e admite fortes pressuposições distribucionais para estimar a distribuição amostral da estatística desejada. Há, no entanto, casos em que essas soluções analíticas não são disponíveis, ou apresentam um grau de complexidade muito grande. Nesse caso, pode-se utilizar, como alternativa, os métodos computacionalmente intensivos, que possibilitam análises livres das suposições sobre os modelos teóricos para a distribuição dos dados e permitem, ainda, o estudo de propriedades amostrais de interesse, mesmo que possuam formas analíticas complexas.

O método “bootstrap” apresentado por Efron (1979), doravante denominado “bootstrap” clássico, é um dos mais importantes métodos computacionalmente intensivos e tem aplicações em inúmeras situações, como apresentam Efron e Tibishirani (1993) e Davison e Hinkley (1997).

A essência do método “bootstrap” é a idéia de que, na ausência de qualquer conhecimento sobre a população, a distribuição dos valores encontrados em uma amostra aleatória de tamanho n seria a melhor forma para obter essa informação. Portanto, para aproximar o que aconteceria se a população fosse reamostrada sucessivamente é razoável fazerem-se reamostragens com reposição da amostra original (MANLY, 1997). Tanto no caso paramétrico quanto no não-paramétrico, pode ser considerado uma ferramenta para obtenção

de estimativas pontuais e intervalares de parâmetros, bem como para a avaliação da precisão de estimativas e testes.

Na sua forma não-paramétrica, consiste na geração de amostras aleatórias do mesmo tamanho da amostra original extraída da população sobre a qual se desejam inferências. Baseia-se, assim, na reprodução do mecanismo probabilístico gerador dos dados originais. Nesse caso a amostragem é feita com reposição, a partir de uma função de distribuição não paramétrica que atribui probabilidades $\frac{1}{n}$ a todas as n observações.

No caso paramétrico, assume-se que a função de distribuição é membro de alguma família paramétrica cujos parâmetros são desconhecidos. Estimam-se, então, esses parâmetros pelo método da máxima verossimilhança e obtém-se a função de distribuição estimada substituindo, na função assumida, os parâmetros pelas suas estimativas. Em seguida, geram-se amostras dessa distribuição estimada.

Baseado em inferência Bayesiana via processo Dirichlet (FERGUSON, 1973), Rubin (1981) apresentou o método “bootstrap” Bayesiano, semelhante ao “bootstrap” clássico. Nesse método, é atribuído, a cada observação, um peso proporcional às probabilidades geradas a partir de uma distribuição a posteriori, considerando-se uma priori imprópria. Essa distribuição a posteriori é centrada em $\frac{1}{n}$, mas varia de uma repetição “bootstrap” para outra. Nesse caso, valores da variável resposta que não são observados têm probabilidade igual a zero, da mesma forma que ocorre sob a função de distribuição amostral.

A distribuição “bootstrap” Bayesiana do parâmetro é semelhante à distribuição “bootstrap” do seu estimador diferindo, no entanto, na forma com que as probabilidades associadas às observações são geradas e na interpretação das distribuições resultantes. Enquanto o “bootstrap” Bayesiano simula uma distribuição a posteriori para o parâmetro, o “bootstrap” clássico simula a distribuição amostral de uma estatística.

2.1.2 Metodologia de superfícies de resposta

As idéias básicas sobre a MSR apareceram no século XIX, aplicadas a estudos que envolviam a relação entre doses de fertilizantes e produtividade (ANDRADE, 1977). Inicialmente, porém, não se discutiu o planejamento específico de experimentos para curvas

e superfícies de resposta, sendo sugerido o uso de planejamentos fatoriais complexos.

O grande impulso na análise de dados experimentais por meio de superfícies de resposta, foi dado pelo artigo de Box e Wilson (1951), no qual os autores desenvolveram métodos de determinação de condições ótimas em pesquisas industriais. Consideraram, basicamente, que a resposta pode ser aproximada, em uma região de interesse, por um polinômio nos níveis de vários fatores ou variáveis envolvidos. Apresentaram, então, planejamentos para a determinação do ponto de resposta ótima, utilizando o menor número possível de pontos, cujos desempenhos foram comparados com os de outros planejamentos experimentais utilizados na época. Foi preconizada, também, a adoção do método de máxima inclinação ascendente na procura de uma região estacionária em torno do ótimo e o estudo da natureza dessa região reduzindo a equação ajustada à sua forma canônica. Nesse trabalho, os autores estavam interessados, principalmente, em problemas de estimação de condições ótimas em trabalhos na área de química, mas relataram que os métodos descritos poderiam ser de grande valia em outros campos onde a experimentação é seqüencial e o erro experimental, razoavelmente pequeno. Esse e vários outros estudos conduzidos por Box e associados, posteriormente, constituíram a base para os trabalhos de pesquisa em MSR. Como exemplos podem ser citados Box e Hunter (1954, 1957), Box e Lucas (1959), Box e Draper (1959, 1963), dentre outros. A literatura existente específica sobre a MSR, por sua vez, é bastante extensa, apresentando diversas aplicações em diferentes campos de pesquisa sendo que revisões sobre o tema são apresentadas por Mead e Pike (1975), Myers; Khuri e Carter Jr. (1989), Myers (1999) e Myers et al. (2004).

Conforme Box e Draper (1987), Khuri e Cornell (1996) e Myers e Montgomery (2002), de uma forma geral, mencionar que uma resposta depende dos níveis de fatores quantitativos é o mesmo que supor a existência de uma função de resposta que represente a verdadeira relação entre estas quantidades. Na prática, não se observa o verdadeiro valor da resposta, mas sim o valor de uma variável aleatória, denominada resposta observada.

Apesar de, em certas ocasiões haver o interesse no modelo global, geralmente desja-se determinar os valores das variáveis preditoras cuja resposta é um máximo ou mínimo, ou seja, procura-se obter o ponto ótimo de operação. Com isso, o pesquisador não precisa de uma modelagem global em toda a região de operacionalidade, mas sim determinar um modelo

empírico que se ajuste à superfície de resposta dentro de uma região de interesse, denominada região experimental. Essa simplificação é importante, pois normalmente, o mecanismo que produz os dados é desconhecido ou pouco compreendido, de modo que a forma matemática da superfície de resposta verdadeira é desconhecida e um modelo empírico é freqüentemente ajustado aos dados. Como modelo empírico, geralmente adota-se um modelo linear nos parâmetros e da forma polinomial, ou nas variáveis preditoras originais ou em alguma transformação dessas. Esta consideração é equivalente a admitir que a verdadeira função de resposta pode ser aproximada localmente pela expansão de sua série de Taylor até o termo de uma certa ordem, na maioria dos casos de segunda ordem (ANDRADE, 1977). Polinômios, além disso, são preferidos devido à facilidade no seu uso e na obtenção de estimadores para os coeficientes do modelo.

Alternativamente, se o mecanismo base (químico, físico, biológico) que produz os dados é conhecido, é possível construir um modelo que o represente chamado modelo mecanístico. Este geralmente contém poucos parâmetros, pode-se ajustar melhor aos dados e permite inferências fora da região experimental considerada, o que não é recomendado no caso de modelos polinomiais. Mead e Pike (1975) discutem algumas dessas aplicações no contexto de ensaios biológicos.

Uma vez ajustado um modelo de superfícies de resposta, muitas indústrias procedem a exploração da região experimental por meio do método de máxima inclinação ascendente, principalmente, nas áreas de melhoria da qualidade e desenvolvimento de novos processos e produtos (MYERS e MONTGOMERY, 2002). Na área de agricultura, no entanto, talvez pela dificuldade de se ter uma experimentação seqüencial, a metodologia tem sido pouco utilizada, exceção feita a experimentos conduzidos em estufas, em ambientes com condições controladas, ou a experimentos de adubação. Nesses casos geralmente o que interessa é obter a produção ou a receita máxima. Além disso, em trabalhos nessa área, costumeiramente estudam-se o desempenho de planos experimentais, o ajuste do modelo da superfície de resposta e a localização e a natureza do ponto estacionário (GOMES, 1989; PIMENTEL GOMES, 1990). Assim, ajustado um modelo de segunda ordem, as coordenadas do ponto estacionário da superfície são facilmente obtidas podendo este ser um ponto de máximo, de mínimo ou de sela. No caso de ser de máximo, no entanto, uma ou mais de suas

coordenadas podem corresponder a um nível negativo de aplicação do nutriente, ou ainda, uma ou mais coordenadas corresponderem a níveis que extrapolem os intervalos usados no experimento. Esta última situação caracteriza uma extrapolação que não deve ser utilizada, principalmente no caso de regressão polinomial. Para evitar essas dificuldades, a MSR deve ser aplicada a ensaios isolados de grande precisão ou a grupos de experimentos conduzidos em condições similares (PIMENTEL GOMES; CONAGIN, 1987).

Campos (1967) trabalhou com 50 ensaios de milho em diferentes locais no estado de São Paulo, individualmente e agrupados de diferentes formas. Concluiu que não se deve aplicar a regressão polinomial a um ensaio individual e que os parâmetros da equação de superfície de resposta, tanto nos ensaios isolados como nos grupos de experimentos, apresentam intervalos de confiança bastante amplos, o que acarreta baixa precisão nas estimativas dos rendimentos.

Num estudo de simulação com 5000 experimentos, Malheiros e Perecin (1983) estudaram os efeitos das doses em relação às coordenadas do ponto de máximo de uma superfície de resposta, visando alcançar maior eficiência na obtenção de pontos de máximo. Sugeriram que certas dificuldades que se têm na análise de ensaios individuais, tendem a desaparecer quando se agrupam ensaios.

Uma das dificuldades levantadas nesses estudos é a ocorrência de pontos de sela e alternativas para superá-las foram apresentadas e comparadas por Sanches (1986) utilizando dados simulados para nove delineamentos experimentais. O autor concluiu que o ponto de máximo fronteira (ponto da fronteira da região estudada pelo experimento em que a equação da superfície apresenta produção máxima) e o ponto de máximo experimental (ponto do delineamento em que ocorre a maior produção) são substitutos do ponto de máximo não encontrado.

Embora até aqui só tenha havido menção à fase de ajuste do modelo empírico e análise dos dados, salienta-se que o planejamento experimental representa uma parte extremamente importante na metodologia de superfície de resposta. A escolha do planejamento consiste, basicamente, na escolha das combinações dos níveis dos diversos fatores e no delineamento dos tratamentos. Essa fase deve ser executada de maneira que resulte na maior aproximação da superfície verdadeira. Também deve permitir a estimativa precisa dos efeitos

de interesse, possibilitar a determinação da variância dessas estimativas e do erro experimental, além da falta de ajuste do modelo.

2.1.3 Regiões de confiança das coordenadas do ponto estacionário

A variação amostral é inerente ao processo experimental e, portanto, é interessante obter uma região de confiança das coordenadas verdadeiras de localização do ponto estacionário para complementar a análise da superfície na qual foi ajustado o polinômio de segunda ordem (BOX; HUNTER, 1954; STABLEIN; CARTER Jr.; WAMPLER, 1983; CARTER Jr. et al., 1986; KHURI; CORNELL, 1996; MYERS; MONTGOMERY, 2002). Dependendo da situação estudada, o ponto estacionário é um máximo, um mínimo ou um ponto de sela. Portanto, a determinação da região de confiança deve ser acompanhada por uma análise canônica para caracterizar a natureza da superfície de resposta.

O conhecimento da região de confiança da localização do ponto estacionário é útil por uma série de razões. Primeiramente, ela dá informações sobre a qualidade da estimativa pontual, por meio da proporção da região experimental abrangida pela região de confiança, indicando se o polinômio tem um ajuste adequado ou não. Nesse caso, quanto menor for essa região maior é a precisão da estimativa.

Uma outra razão, apontada em Carter Jr. et al. (1986), no contexto da indústria farmacêutica, é que ela fornece informação sobre sinergismo terapêutico, um problema comum que envolve tempo de sobrevivência em experimentos nessa área. Segundo os autores, se a combinação ótima da dosagem exclui todas as combinações individuais com dosagem zero, então existe evidência de que todos os componentes da droga são benéficos para a terapia, considerando-se uma forma sinérgica de ação. Esse conceito pode ser aplicado, por extensão, a experimentos para a formulação de produtos, nos quais deseja-se saber se um ingrediente deve, ou não, ser incluído na formulação (CASTILLO; CAHYA, 2001). Regiões de confiança também são úteis no caso de situações envolvendo respostas múltiplas.

Para determinar-se a região de confiança da localização do ponto estacionário têm sido sugeridas várias alternativas, sendo a primeira delas, a de Box e Hunter (1954), baseada na pressuposição de normalidade dos erros aleatórios do modelo. Stablein; Carter

Jr.e Wampler (1983), Carter Jr. et al. (1986), por sua vez, estenderam a idéia de Box e Hunter a casos de superfícies em que há restrições que limitam a região experimental, como, por exemplo, experimentos com misturas, utilizando multiplicadores de Lagrange para incorporar restrições à análise. Esse método, no entanto, ignora a variabilidade amostral dos multiplicadores; suposição que não é válida porque os multiplicadores são dependentes das estimativas dos parâmetros do modelo e essas estimativas são sujeitas à variabilidade amostral. Em vista disso, Peterson; Cahya e Castillo (2002) apresentaram uma abordagem generalizada que não usa os multiplicadores de Lagrange evitando a dificuldade técnica de se incorporar a sua variabilidade amostral na análise. Cahya; Castillo e Peterson (2004), por sua vez, propuseram algumas modificações ao método anterior, tornando o algoritmo mais rápido e mais acurado.

Recentemente, foram introduzidas formas diferentes para a obtenção da região de confiança baseadas na metodologia “bootstrap” não paramétrica, descrita em Efron (1979, 1982), Efron e Tibshirani (1993) e Davison e Hinkley (1997). Segundo Freitas (1993) essa técnica é uma ótima ferramenta para a construção da região de confiança para a localização do ponto de resposta ótimo em exploração de superfícies de resposta. Para definir a referida região, a autora baseou-se na distância de Mahalanobis entre as coordenadas do ponto estacionário da amostra observada e aquelas obtidas através das estimativas “bootstrap”. Zocchi e Silva (2003) e Silva e Zocchi (2004), também usaram essa técnica mas, desta vez, aliada a métodos não paramétricos de estimação de funções densidade de probabilidade (SILVERMAN, 1986), para a construção dessas regiões de confiança. Segundo os autores, a metodologia apresentada tem implementação bastante simples e possui a vantagem de não requerer a pressuposição de normalidade.

Quando, no entanto, trabalha-se com delineamentos em que há poucas repetições dos tratamentos envolvidos ou que são quase saturados ou saturados, a metodologia bootstrap padrão pode não propiciar uma boa cobertura do espaço paramétrico. Como alternativa, propõe-se utilizar o método “bootstrap” Bayesiano que produz uma distribuição mais suave, propiciando uma melhor cobertura no espaço paramétrico. Isso é consequência da forma com que as probabilidades são atribuídas às observações, discretas no caso do primeiro método e contínuas para o segundo.

2.2 Metodologia

2.2.1 Material

Os dados utilizados para ilustrar o método proposto são oriundos de um estudo com fertilizantes descrito em Khuri e Cornell (1996). Este foi conduzido com o objetivo de observar o efeito de dois tipos de fertilizantes, uma formulação NPK padrão (NPK) e um suplemento nutricional (SN), sobre a produtividade de amendoim. Os tratamentos consistiram nas diferentes combinações dos níveis dos dois fertilizantes e foram atribuídos às parcelas (unidades experimentais) de acordo com um delineamento composto central rotacional, com duas repetições. Os resultados são apresentados na Tabela 1.

2.2.2 Métodos

Nesta seção é apresentada a metodologia para obtenção das regiões de confiança com pressuposição de normalidade dos erros do modelo (BOX e HUNTER, 1954), as baseadas na técnica “bootstrap” (FREITAS, 1993; ZOCCHI e SILVA, 2003; SILVA E ZOCCHI, 2004) e a metodologia proposta neste trabalho baseada no método “bootstrap” Bayesiano. Também são descritos os procedimentos adotados nos estudos de simulação.

2.2.2.1 Regiões de confiança

Seja Y uma variável resposta e $\mathbf{x} = [x_1, \dots, x_k]^T$ o vetor de k variáveis explicativas. Considere o modelo de superfície de resposta de segundo grau completo dado por:

$$Y = \beta_0 + \mathbf{x}^T \mathbf{b} + \mathbf{x}^T \mathbf{B} \mathbf{x} + \epsilon \quad (1)$$

em que, β_0 é um parâmetro; $\mathbf{b} = [\beta_1, \dots, \beta_k]^T$ é o vetor de parâmetros de primeira ordem; $\mathbf{B}$ é a matriz de parâmetros de segunda ordem, de dimensão $k \times k$, dada por:

Tabela 1 - Produtividade de amendoim (libra/parcela) em função das quantidades (libra/parcela) de dois fertilizantes, NPK e SN. (rep.=repetição)

<i>NPK</i>	<i>SN</i>	Produtividade	
		rep. 1	rep. 2
50,0	15,0	7,52	8,12
120,0	15,0	12,37	11,84
50,0	25,0	13,55	12,35
120,0	25,0	16,48	15,32
35,5	20,0	8,63	9,44
134,5	20,0	14,22	12,57
85,0	12,9	7,90	7,33
85,0	27,1	16,49	17,40
85,0	20,0	15,73	17,00

$$\mathbf{B} = \begin{bmatrix} \beta_{11} & \frac{\beta_{12}}{2} & \dots & \frac{\beta_{1k}}{2} \\ \frac{\beta_{12}}{2} & \beta_{22} & \dots & \frac{\beta_{2k}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\beta_{1k}}{2} & \frac{\beta_{2k}}{2} & \dots & \beta_{kk} \end{bmatrix}$$

e ϵ é o erro aleatório.

Dados $\mathbf{b}$ e $\mathbf{B}$, a localização $\boldsymbol{\theta}$ do ponto estacionário é o vetor $\mathbf{x}$ solução do sistema de equações,

$$\frac{\partial Y}{\partial \mathbf{x}} = \mathbf{b} + 2\mathbf{B}\mathbf{x} = \mathbf{0}$$

em que $\mathbf{0}$ é um vetor de zeros de dimensão k , ou seja,

$$\boldsymbol{\theta} = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} ,$$

desde que $\mathbf{B}$ seja não singular.

Os sinais dos autovalores de $\mathbf{B}$ fornecem informação sobre a natureza do ponto estacionário $\boldsymbol{\theta}$ da seguinte forma: se todos os autovalores são menores do que zero, tem-se que $\boldsymbol{\theta}$ é um ponto de máximo; se todos são maiores do que zero, $\boldsymbol{\theta}$ é um ponto de mínimo e se existem autovalores maiores do que zero e outros menores do que zero tem-se que $\boldsymbol{\theta}$ é um ponto de sela.

O primeiro método de construção de uma região de confiança sobre a locação do ponto estacionário foi apresentado por Box e Hunter (1954), descrito a seguir.

Considere um conjunto de n observações da variável resposta Y , $\{y_1, \dots, y_n\}$, a $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$. Sejam, $\hat{\mathbf{B}}$ e $\hat{\mathbf{b}}$ a matriz e o vetor de estimativas de mínimos quadrados dos parâmetros de $\mathbf{B}$ e $\mathbf{b}$, ou seja, $\hat{\mathbf{b}} = [\hat{\beta}_1, \dots, \hat{\beta}_k]^T$ e

$$\hat{\mathbf{B}} = \begin{bmatrix} \hat{\beta}_{11} & \frac{\hat{\beta}_{12}}{2} & \dots & \frac{\hat{\beta}_{1k}}{2} \\ \frac{\hat{\beta}_{12}}{2} & \hat{\beta}_{22} & \dots & \frac{\hat{\beta}_{2k}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\hat{\beta}_{1k}}{2} & \frac{\hat{\beta}_{2k}}{2} & \dots & \hat{\beta}_{kk} \end{bmatrix}.$$

Seja, ainda,

$$\boldsymbol{\delta} = \left. \frac{\partial Y}{\partial \mathbf{x}} \right|_{\mathbf{b}=\hat{\mathbf{b}}, \mathbf{B}=\hat{\mathbf{B}}, \mathbf{x}=\boldsymbol{\theta}} = \hat{\mathbf{b}} + 2\hat{\mathbf{B}}\boldsymbol{\theta}$$

e

$$\hat{\boldsymbol{\theta}} = -\frac{1}{2}\hat{\mathbf{B}}^{-1}\hat{\mathbf{b}},$$

o estimador de mínimos quadrados de $\boldsymbol{\theta}$.

Supondo-se que $\epsilon_i \sim N(0, \sigma^2)$, $\{i = 1, \dots, n\}$ e que os erros são independentes, tem-se que

$$\begin{aligned} E(\boldsymbol{\delta}) &= E(\hat{\mathbf{b}} + 2\hat{\mathbf{B}}\boldsymbol{\theta}) = E(\hat{\mathbf{b}}) + 2E(\hat{\mathbf{B}})\boldsymbol{\theta} \\ &= \mathbf{b} + 2\mathbf{B}\left(-\frac{1}{2}\right)\mathbf{B}^{-1}\mathbf{b} = \mathbf{0} \end{aligned} \quad (2)$$

e

$$Var(\boldsymbol{\delta}) = \mathbf{V}\sigma^2,$$

em que $\mathbf{V}$ é a matriz de variâncias e covariâncias de $\boldsymbol{\delta}$. Conseqüentemente, tem-se que $\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{V}\sigma^2)$.

A matriz $\mathbf{V}$ é facilmente obtida tomando-se como base resultados apresentados em Böckenholt (1989) e Gilmour e Draper (2003). Considere o modelo (1) escrito como:

$$Y = \mathbf{z}^T \boldsymbol{\beta}, \quad (3)$$

em que,

$$\mathbf{z} = [1, x_1, \dots, x_k, x_1^2, \dots, x_k^2, x_1x_2, \dots, x_{(k-1)}x_k]^T$$

e $\boldsymbol{\beta} = [\beta_0, \beta_1, \dots, \beta_k, \beta_{11}, \dots, \beta_{kk}, \dots, \beta_{12}, \dots, \beta_{(k-1)k}]^T$.

Seja, ainda,

$$\mathbf{C}(\boldsymbol{\theta}) = \left. \frac{\partial \mathbf{z}^T}{\partial \mathbf{x}} \right|_{\mathbf{x}=\boldsymbol{\theta}}$$

o valor de $\mathbf{C}(\mathbf{x})$ obtido em $\boldsymbol{\theta}$. Então, pode-se escrever $\mathbf{V}$ da seguinte forma:

$$\mathbf{V} = \mathbf{C}(\boldsymbol{\theta})(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{C}(\boldsymbol{\theta})^T$$

em que

$$\mathbf{Z} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} & x_{11}^2 & \cdots & x_{1k}^2 & x_{11}x_{12} & \cdots & x_{1(k-1)}x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} & x_{21}^2 & \cdots & x_{2k}^2 & x_{21}x_{22} & \cdots & x_{2(k-1)}x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} & x_{n1}^2 & \cdots & x_{nk}^2 & x_{n1}x_{n2} & \cdots & x_{n(k-1)}x_{nk} \end{bmatrix}$$

Como caso particular tem-se que, para $k = 2$,

$$\mathbf{V} = \begin{bmatrix} Var(\hat{\beta}_1) + 4Var(\hat{\beta}_{11})\theta_1^2 + Var(\hat{\beta}_{12})\theta_2^2 & Var(\hat{\beta}_{12})\theta_1\theta_2 + 4Cov(\hat{\beta}_{11}, \hat{\beta}_{22})\theta_1\theta_2 \\ Var(\hat{\beta}_{12})\theta_1\theta_2 + 4Cov(\hat{\beta}_{11}, \hat{\beta}_{22})\theta_1\theta_2 & Var(\hat{\beta}_2) + Var(\hat{\beta}_{12})\theta_1^2 + 4Var(\hat{\beta}_{22})\theta_2^2 \end{bmatrix} \sigma^2.$$

Nessas condições a região de $100(1 - \alpha)\%$ de confiança para $\boldsymbol{\theta}$ pode ser definida como o conjunto de valores de $\boldsymbol{\theta}$ que satisfazem a desigualdade:

$$\frac{\boldsymbol{\delta}^T \mathbf{V}^{-1} \boldsymbol{\delta}}{2\hat{\sigma}^2} \leq F_{\alpha, k, \nu}, \quad (4)$$

em que $\hat{\sigma}^2$ é o quadrado médio residual da análise da variância, com ν graus de liberdade e $F_{\alpha, k, \nu}$ é o percentil de ordem $100(1 - \alpha)\%$ da distribuição F com k e ν graus de liberdade.

Note que, embora a equação (4) tenha uma forma explícita, determinar a região de confiança de Box e Hunter pode ser uma tarefa algébrica difícil, sendo que há, ainda, a possibilidade de tal método não ser adequado quando os erros não possuírem distribuição normal.

Alternativamente, pode-se utilizar a metodologia proposta por Zocchi e Silva (2003) e estendida por Silva e Zocchi (2004), baseada na técnica “bootstrap”, cujo procedimento será apresentado a seguir.

Considere o conjunto de observações $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$, obtenha $\hat{\beta}_0$, $\hat{\mathbf{b}}$ e $\hat{\mathbf{B}}$, e o conjunto de resíduos padronizados $\{\zeta_1, \dots, \zeta_n\}$, em que

$$\zeta_i = \left[y_i - (\hat{\beta}_0 + \mathbf{x}_i^T \hat{\mathbf{b}} + \mathbf{x}_i^T \hat{\mathbf{B}} \mathbf{x}_i) \right] \left(\frac{n}{n-p} \right)^{\frac{1}{2}}, \quad (i = 1, \dots, n)$$

e p é o número total de parâmetros do modelo, dado por $p = (k + 1)(k + 2)/2$.

Em seguida, retire M (geralmente, $M > 1000$) amostras de tamanho n , com reposição, do conjunto de resíduos $\{\zeta_1, \dots, \zeta_n\}$, representadas por $\{\zeta_1^{*(m)}, \dots, \zeta_n^{*(m)}\}$, ($m = 1, \dots, M$). Para cada uma dessas amostras, denominadas amostras “bootstrap”, construa um novo conjunto de dados:

$$\{(\mathbf{x}_1, \hat{\beta}_0 + \mathbf{x}_1^T \hat{\mathbf{b}} + \mathbf{x}_1^T \hat{\mathbf{B}} \mathbf{x}_1 + \zeta_1^{*(m)}), \dots, (\mathbf{x}_n, \hat{\beta}_0 + \mathbf{x}_n^T \hat{\mathbf{b}} + \mathbf{x}_n^T \hat{\mathbf{B}} \mathbf{x}_n + \zeta_n^{*(m)})\}, \quad (m = 1, \dots, M).$$

Para cada um desses M conjuntos estime β_0 , $\mathbf{b}$, $\mathbf{B}$ e $\boldsymbol{\theta}$, obtendo dessa forma, $\{\hat{\boldsymbol{\theta}}^{*(1)}, \dots, \hat{\boldsymbol{\theta}}^{*(M)}\}$, em que $\hat{\boldsymbol{\theta}}^{*(m)}$ é a estimativa de $\boldsymbol{\theta}$ referente à m -ésima amostra “bootstrap”.

Baseando-se em $\{\hat{\boldsymbol{\theta}}^{*(1)}, \dots, \hat{\boldsymbol{\theta}}^{*(M)}\}$, estima-se a função densidade de probabilidade conjunta de $\boldsymbol{\theta} = [\theta_1, \dots, \theta_k]^T$, por meio de

$$f(\boldsymbol{\theta}) = f(\theta_1, \dots, \theta_k) = \frac{1}{M \prod_{j=1}^k h_j} \sum_{m=1}^M \varphi\left(\frac{\theta_1 - \hat{\theta}_1^{*(m)}}{h_1}\right) \cdots \varphi\left(\frac{\theta_k - \hat{\theta}_k^{*(m)}}{h_k}\right).$$

em que, h_j ($j = 1, \dots, k$) são as constantes de suavização e φ é a função densidade kernel, no caso, normal padrão, dada por $\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}t^2\}$, para $t \in \Re$. Seguindo a sugestão de Bowman e Azzalini (1997), utiliza-se, como constantes de suavização,

$$h_j = \left[\frac{4}{M(k+2)} \right]^{\frac{1}{(k+4)}} \sqrt{s_{jj}}$$

em que,

$$s_{jj} = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}_j^{*(m)} - \bar{\theta}_j^*)^2$$

e

$$\bar{\theta}_j^* = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_j^{*(m)}.$$

Em seguida, calcule o valor de $f(\hat{\boldsymbol{\theta}}^{*(m)})$, para cada m ($m = 1, \dots, M$). Seja P_α o percentil de ordem α do conjunto de valores $\{f(\hat{\boldsymbol{\theta}}^{*(1)}), \dots, f(\hat{\boldsymbol{\theta}}^{*(M)})\}$. A região de $100 \times (1 - \alpha)\%$ de confiança para $\boldsymbol{\theta}$ será dada por:

$$(\boldsymbol{\theta}) \in \Re^k : f(\boldsymbol{\theta}) \leq P_\alpha.$$

Sejam

$$\hat{\boldsymbol{\theta}}^* = \left[\hat{\theta}_1^{*(m)}, \hat{\theta}_2^{*(m)}, \dots, \hat{\theta}_k^{*(m)} \right]^T,$$

$$\bar{\boldsymbol{\theta}}^* = \left[\bar{\theta}_1^*, \bar{\theta}_2^*, \dots, \bar{\theta}_k^* \right]^T,$$

$$\boldsymbol{\Sigma}_r = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{12} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1k} & s_{2k} & \cdots & s_{kk} \end{bmatrix} e$$

$$s_{jj'} = \frac{1}{M} \sum_{m=1}^M (\hat{\theta}_j^{*(m)} - \bar{\theta}_j^*)(\hat{\theta}_{j'}^{*(m)} - \bar{\theta}_{j'}^*).$$

Uma forma alternativa ao uso de densidades não paramétricas para a obtenção das regiões de confiança para as coordenadas do ponto estacionário, é defini-las por meio da distância de Mahalanobis (FREITAS, 1993). Seja $DM(\boldsymbol{\theta})$ a distância de Mahalanobis do vetor $\boldsymbol{\theta}$ ao vetor de médias “bootstrap” $\bar{\boldsymbol{\theta}}^*$ dada por,

$$DM(\boldsymbol{\theta}) = (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*)^T \boldsymbol{\Sigma}_r^{-1} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*)$$

e

$$DM(\boldsymbol{\theta}^{*(m)}) = (\boldsymbol{\theta}^{*(m)} - \bar{\boldsymbol{\theta}}^*)^T \boldsymbol{\Sigma}_r^{-1} (\boldsymbol{\theta}^{*(m)} - \bar{\boldsymbol{\theta}}^*)$$

a distância de Mahalanobis de $\boldsymbol{\theta}^{*(m)}$ a $\bar{\boldsymbol{\theta}}^*$, ($m = 1, \dots, M$).

Seja, ainda, Q o percentil de ordem α do conjunto de distâncias $\{DM(\hat{\boldsymbol{\theta}}^{*(1)}), DM(\hat{\boldsymbol{\theta}}^{*(2)}), \dots, DM(\hat{\boldsymbol{\theta}}^{*(M)})\}$. A região de $100 \times (1 - \alpha)\%$ de confiança para $\boldsymbol{\theta}$ será dada por:

$$\boldsymbol{\theta} \in \mathfrak{R}^k : DM(\boldsymbol{\theta}) \leq Q_\alpha.$$

Uma outra maneira de se obter uma amostra “bootstrap”, descrita por Efron e Tibshirani (1993) e que é menos sensível às pressuposições do modelo, é a de se reamostrarem os pares de observações $(\mathbf{x}_i, y_i)$. Assim, a partir do conjunto de dados $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ são retiradas M amostras de tamanho n , com reposição, $\{(\mathbf{x}_1, y_1)^{*(m)}, \dots, (\mathbf{x}_n, y_n)^{*(m)}\}$, ($m = 1, \dots, M$) e, a partir delas estima-se β_0 , $\mathbf{b}$, $\mathbf{B}$ e $\boldsymbol{\theta}$, obtendo dessa forma, $\{\hat{\boldsymbol{\theta}}^{*(1)}, \dots, \hat{\boldsymbol{\theta}}^{*(M)}\}$, em que $\hat{\boldsymbol{\theta}}^{*(m)}$ é a estimativa de $\boldsymbol{\theta}$ referente à m -ésima amostra “bootstrap”. Essa forma alternativa de obter uma amostra “bootstrap” é, no entanto, recomendada quando há erros de medida em $\mathbf{x}$, o que não é o caso em questão. Além disso, a matriz do delineamento baseada na amostra “bootstrap” pode ser singular, especialmente para delineamentos quase saturados.

Um método similar ao “bootstrap” clássico é aquele apresentado por Rubin (1981), baseado em inferência Bayesiana via processo Dirichlet (FERGUSON, 1973). Segundo esse método, atribui-se a cada observação um determinado peso igual às probabilidades geradas de uma distribuição a posteriori, obtida com uma priori imprópria. Essa distribuição a posteriori é centrada em $\frac{1}{n}$, mas varia de uma repetição “bootstrap” para outra. Valores da variável que não são observados têm probabilidade igual a zero, da mesma forma que ocorre sob a função de distribuição amostral.

Para apresentar o método proposto por Rubin (1981) vamos, inicialmente, considerar a lógica do método “bootstrap”. Seja Ω um espaço amostral discreto, $\Omega = \{1, 2, \dots, L\}$. O fato de se assumir um espaço amostral discreto não constitui uma restrição ao uso da metodologia “bootstrap” (EFRON, 1982), uma vez que se pode tornar o número L de categorias discretas, arbitrariamente grande. Se o espaço for a reta real, pode-se, por exemplo, dividir o intervalo $[-10^{10}, 10^{10}]$ em 2×10^{20} intervalos de comprimento 10^{-10} . Então, $L = 2 \times 10^{20} + 2$, contando os intervalos semi-infinitos finais, e para propósitos práticos a discretização não afetará as inferências.

Seja $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$, uma amostra aleatória simples de uma população com função de distribuição F , e $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, os seus respectivos valores. Tendo-se uma função de distribuição empírica ou amostral (fda), $\hat{F}$, com probabilidade $\frac{1}{n}$ para cada uma das observações $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, pode-se expressar F como $\mathbf{f} = (f_1, f_2, \dots, f_L)$, em que $f_l = \text{Prob}[(\mathbf{x}, Y) = l]$, e $\hat{F}$ como $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_L)$, em que $\hat{f}_l = \frac{\#[(\mathbf{x}_i, y_i) = l]}{n}$, com $\sum_{l=1}^L \hat{f}_l = 1$. Assim, a distribuição de $\hat{f}_l$ é proporcional à distribuição binomial, $Bi(n, f_l)$.

$$(\hat{f}_l | n; f_l) \propto \frac{1}{n} Bi(n, f_l).$$

Considerando que cada tentativa é independente e que qualquer um dos n elementos amostrais tem a mesma probabilidade, $\frac{1}{n}$, de ser selecionado, a distribuição de $\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \dots, \hat{f}_L)$ é proporcional à distribuição multinomial, $Mn_L(n, \mathbf{f})$.

$$(\hat{\mathbf{f}} | n; \mathbf{f}) \propto \frac{1}{n} Mn_L(n, \mathbf{f}) \quad (5)$$

A distribuição “bootstrap” de $(\mathbf{x}_i, Y_i)^* \stackrel{i.i.d.}{\sim} \hat{F}$ ($i = 1, \dots, n$), pode ser descrita

em termos de $\hat{\mathbf{f}}^* = (\hat{f}_1^*, \hat{f}_2^*, \dots, \hat{f}_L^*)$, em que $\hat{f}_l^* = \frac{nr \cdot [(\mathbf{x}_i, y_i)^* \in l]}{n}$, com $\sum_{l=1}^L \hat{f}_l^* = 1$. Ela é a mesma como em (5), com $\hat{\mathbf{f}}$ no lugar de $\mathbf{f}$,

$$(\hat{\mathbf{f}}^* | n; \hat{\mathbf{f}}) \propto \frac{1}{n} Mn_L(n, \hat{\mathbf{f}}) \quad (6)$$

Portanto, o conceito chave do método é que cada repetição “bootstrap” cria uma coleção de pesos, proporcional a $Mn_L(n, \hat{\mathbf{f}})$, que representa a proporção de vezes que cada par $(\mathbf{x}_i, y_i)$ foi amostrado, gerando, portanto uma verossimilhança para os dados.

A idéia de Rubin (1981) foi, a partir dessa verossimilhança e de uma distribuição a priori conjugada imprópria, obter uma coleção dos pesos amostrando a partir da posteriori.

Considere como a priori sobre $\mathbf{f}$ uma distribuição Dirichlet, conforme descrita em Wilks (1962), com parâmetro a , $\mathbf{f} \sim Di_L(a, a, \dots, a)$, com função densidade dada por:

$$p(\mathbf{f}) = \frac{\Gamma(a + a + \dots + a)}{\prod_{l=1}^L \Gamma(a)} \prod_{l=1}^L f_l^{a-1}.$$

Tendo-se observado $\hat{\mathbf{f}}$, a distribuição a posteriori de $\mathbf{f}$, é:

$$\mathbf{f} | \hat{\mathbf{f}} \sim Di_L(a + n\hat{f}_1, a + n\hat{f}_2, \dots, a + n\hat{f}_L).$$

com função densidade proporcional a $\prod_{l=1}^L f_l^{n\hat{f}_l + a - 1}$. Fazendo-se $a = 0$ para representar a informação a priori, resulta em $\mathbf{f} | \hat{\mathbf{f}} \sim Di_L(n\hat{f}_1, n\hat{f}_2, \dots, n\hat{f}_L)$, cuja função densidade é:

$$p(\mathbf{f} | \hat{\mathbf{f}}) = \frac{\Gamma(n\hat{f}_1 + n\hat{f}_2 + \dots + n\hat{f}_L)}{\prod_{l=1}^L \Gamma(n\hat{f}_l)} \prod_{l=1}^L f_l^{n\hat{f}_l - 1}. \quad (7)$$

Assim, no “bootstrap” clássico, os pesos atribuídos aos pares $(\mathbf{x}_i, y_i)$ têm distribuição proporcional a $Mn_L(n, \hat{\mathbf{f}})$, e, no “bootstrap” Bayesiano esses pesos apresentam distribuição $Di_L(1, 1, \dots, 1)$. Essa duas distribuições são bastante similares, sendo que,

1. Ambas têm suporte sobre aquelas categorias com $\hat{f}_l > 0$, isto é, sobre aquelas categorias nas quais os dados foram observados;
2. Ambas têm vetor de esperanças igual a $\hat{\mathbf{f}}$;
3. As matrizes de variâncias e covariâncias são aproximadamente iguais, $Cov(\mathbf{f}|\hat{\mathbf{f}}) = \frac{\Sigma_{\hat{f}}}{n+1}$ e $Cov_*(\hat{\mathbf{f}}^*|\hat{\mathbf{f}}) = \frac{\Sigma_{\hat{f}}}{n}$, em que $\Sigma_{\hat{f}}$ tem elementos iguais a $\hat{f}_l(1 - \hat{f}_l)$ na diagonal principal e $-\hat{f}_l\hat{f}_{l'}$ fora da diagonal principal;
4. As matrizes de correlação são iguais com os elementos fora da diagonal principal iguais a $\frac{-\hat{f}_l\hat{f}_{l'}}{\sqrt{(\hat{f}_l(1-\hat{f}_l))(\hat{f}_{l'}(1-\hat{f}_{l'}))}}$.

As características indicadas das duas distribuições são facilmente verificadas.

Conforme se demonstra, de (6) tem-se:

$$E(\hat{f}_l^*|\hat{f}_l) = \frac{1}{n} n \hat{f}_l = \hat{f}_l;$$

$$Var(\hat{f}_l^*|\hat{f}_l) = \frac{1}{n^2} n \hat{f}_l (1 - \hat{f}_l) = \frac{\hat{f}_l (1 - \hat{f}_l)}{n}$$

e

$$Cov(\hat{f}_l^*, \hat{f}_{l'}^*|\hat{f}_l, \hat{f}_{l'}) = \frac{1}{n^2} (-n \hat{f}_l \hat{f}_{l'}) = \frac{-\hat{f}_l \hat{f}_{l'}}{n}.$$

E, de (7) tem-se:

$$E(f_l|\hat{f}_l) = \frac{n\hat{f}_l}{n\hat{f}_l + n\hat{f}_2 + \dots + n\hat{f}_L} = \hat{f}_l;$$

$$Var(f_l|\hat{f}_l) = \frac{n\hat{f}_l(n - n\hat{f}_l)}{n^2(n + 1)} = \frac{\hat{f}_l(1 - \hat{f}_l)}{n + 1}$$

e

$$Cov(f_l, f_{l'}|\hat{f}_l, \hat{f}_{l'}) = \frac{-(n\hat{f}_l n\hat{f}_{l'})}{n^2(n + 1)} = \frac{-\hat{f}_l \hat{f}_{l'}}{n + 1}.$$

A distribuição Dirichlet, com as características desejadas, pode ser facilmente simulada, considerando-se os seguintes resultados apresentados em Wilks (1962):

1. Seja X uma variável aleatória com função densidade f e função de distribuição F . Seja $Y = F(X)$. Então, Y será uniformemente distribuída no intervalo $[0, 1]$.
2. Sejam $x_{(1)}, \dots, x_{(n)}$ as estatísticas de ordem de uma amostra extraída de uma função de distribuição contínua $F(x)$. Então, as diferenças $g_1 = F(x_1), g_2 = F(x_2) - F(x_1), \dots, g_n = F(x_n) - F(x_{n-1})$, são variáveis aleatórias com a distribuição Dirichlet, com parâmetros $Di_{(n-1)}(1, 1, \dots, 1)$.

Para ter-se uma idéia do grau de similaridade mencionado, considere ϕ um parâmetro sobre o qual se desejam inferências. Suponha que W é uma variável binária (0 ou 1), em que $\phi = Pr(W = 1)$ é o parâmetro de interesse. Seja n_1 o número de valores de $w_i = 1$ e $n - n_1$ o número de valores de $w_i = 0$. Para cada repetição do “bootstrap” Bayesiano, obtém-se o vetor de probabilidades, $\mathbf{g} = (g_1, g_2, \dots, g_n)$, a partir das diferenças das estatísticas de ordem da distribuição uniforme $U(0, 1)$. A seguir, multiplica-se $\mathbf{g}$ pelo vetor dos n valores observados de W . Assim, haverá n_1 probabilidades associadas a $w_i = 1$ e $n - n_1$ a $w_i = 0$. Se denominarmos a soma dessas n_1 probabilidades de S_1 , o valor de ϕ em cada amostra bootstrap será igual a S_1 . A repetição do processo gera a distribuição “bootstrap” Bayesiana de ϕ . Dado que, em cada repetição, tem-se $n - 1$ variáveis i.i.d. $U(0, 1)$, as diferenças entre as estatísticas de ordem têm distribuição $Di_{n-1}(1, 1, \dots, 1)$. De acordo com um resultado apresentado em Wilks (1962), a soma de d diferenças entre as estatísticas de ordem de uma função de distribuição contínua tem distribuição Beta, $Be(d, n - d + 1)$. Portanto, a soma de n_1 dessas diferenças, S_1 , segue a distribuição Beta, $Be(n_1, n - n_1)$ e, conseqüentemente, a distribuição “bootstrap” Bayesiana de ϕ é $Be(n_1, n - n_1)$.

Considerando-se que cada w_i na amostra segue a distribuição Bernoulli com parâmetro ϕ a verossimilhança é igual a $\phi^{n_1}(1 - \phi)^{n - n_1}$. Assim, a distribuição a priori de ϕ é proporcional a $[\phi(1 - \phi)]^{-1}$.

Observa-se que a distribuição bootstrap de $\hat{\phi}^*$, a proporção amostrada de valores de $w_i = 1$, é $\frac{Bin(n; \frac{n_1}{n})}{n}$, enquanto que a distribuição de ϕ é Beta, $Be(n_1, n - n_1)$. Assim, ϕ e $\hat{\phi}^*$ tem médias iguais a $\frac{n_1}{n}$ mas, $\hat{\phi}^*$ tem variância igual a $\frac{n_1(n - n_1)}{n^3}$ e ϕ tem variância igual a $\frac{n_1(n - n_1)}{n^2(n + 1)}$.

Para se implementar o método “bootstrap” Bayesiano, considere o conjunto de observações $(\mathbf{x}_1, Y_1), \dots, (\mathbf{x}_n, Y_n)$. Os passos são os seguintes:

1. Gere uma amostra $(u_1, u_2, \dots, u_{n-1})$ de tamanho $n - 1$ de uma distribuição uniforme no intervalo $[0, 1]$;
2. Disponha os valores gerados em ordem crescente. Sejam $u_{(1)}, u_{(2)}, \dots, u_{(n-1)}$ esses valores. Considere $u_{(0)} = 0$ e $u_{(n)} = 1$;
3. Determine $g_i = u_{(i)} - u_{(i-1)}$ com $i = 1, 2, \dots, n$. Seja $\mathbf{g} = [g_1, \dots, g_n]$ e $\mathbf{G} = \text{diag}(\mathbf{g})$;
4. Determine o vetor de parâmetros do modelo:

$$\boldsymbol{\beta} = (\mathbf{Z}^T \mathbf{G} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{G} \mathbf{Y}$$

5. Obtenha, a partir do item anterior, β_0 , $\mathbf{b}$, $\mathbf{B}$ e $\boldsymbol{\theta}$;
6. Repita M vezes os passos 1, 2, 3, 4 e 5.

De posse da distribuição “bootstrap” Bayesiana de $\boldsymbol{\theta}$, a obtenção das regiões de confiança foi efetuada utilizando-se os métodos de cálculo da distância de Mahalanobis e por meio do método não-paramétrico de estimação de densidade, conforme descrito anteriormente.

2.2.2.2 Simulação

A fim de avaliar a metodologia proposta neste trabalho, foram construídos, por meio de simulação de Monte Carlo, 1000 conjuntos de dados. Considerou-se um modelo polinomial de segundo grau completo e um delineamento composto central com duas variáveis preditoras, conforme descrito a seguir.

Seja $k = 2$ e o modelo (1) escrito na sua forma canônica, ou seja,

$$Y = Y_{\theta} + (\mathbf{x} - \boldsymbol{\theta})^T \mathbf{O} \boldsymbol{\Lambda} \mathbf{O}^T (\mathbf{x} - \boldsymbol{\theta}) + \epsilon \quad (8)$$

em que, Y_θ é o valor da variável resposta para $\mathbf{x} = \boldsymbol{\theta}$, $\mathbf{O} = [\mathbf{o}_1 \ \mathbf{o}_2]$ a matriz de dimensões 2×2 formada pelos autovalores normalizados $\mathbf{o}_1$ e $\mathbf{o}_2$ associados aos autovalores λ_1 e λ_2 da matriz $\mathbf{B}$ e

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Estipulou-se, para efeito de simulação o valor de $Y_\theta = 100$, $\boldsymbol{\theta} = \begin{bmatrix} 0,2 \\ 0,4 \end{bmatrix}$, de forma que as coordenadas do ponto estacionário pertençam à região experimental, e

$$\mathbf{O} = \begin{bmatrix} \cos 30^\circ & -\sin 30^\circ \\ \sin 30^\circ & \cos 30^\circ \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sqrt{3} & -1 \\ 1 & \sqrt{3} \end{bmatrix}.$$

Considerou-se ainda ϵ_i com distribuições Normal $N(0; 0, 5)$, $N(0; 0, 8)$, $N(0, 1)$ e Lognormal $LN(0, 1)$. Para essa última distribuição os valores da média e do desvio padrão são obtidos a partir dos valores em escala logarítmica. Para cada uma dessas situações foram escolhidos os valores de $(\lambda_1; \lambda_2)$ dados por: $(-0,025; -0,017)$, $(-0,25; -0,17)$, $(-0,64; -0,59)$ ou $(-1,25; -1,17)$, de forma a obter-se superfícies de resposta mais planas (autovalores com valores absolutos menores), ou com maior grau de curvatura (autovalores com maiores valores absolutos). Em cada uma das situações indicadas foram geradas amostras “bootstrap” de tamanho $N = 1000$. A seguir, avaliou-se, a porcentagem de vezes em que a região de $(1 - \alpha)100\%$ de confiança cobriu o ponto estacionário verdadeiro, considerando-se os níveis de significância $\alpha = 0,05$ e $\alpha = 0,10$, pelos métodos de Box e Hunter (1954), distância de Mahalanobis e contornos de densidade não paramétrica bivariada. As porcentagens de cobertura foram analisadas utilizando-se os quantis de uma distribuição binomial de parâmetros 0,95 e 0,90 e um número de repetições igual a 1000. Este procedimento baseia-se no fato de que em cada conjunto de dados simulado, a região de confiança contém ou não o verdadeiro valor verdadeiro do parâmetro, e considerando-se que o número de conjuntos de dados simulados é fixo e que eles são gerados independentemente.

2.2.2.3 Análise dos dados de produção de amendoim

A metodologia proposta foi aplicada ao conjunto de dados de produção amendoim apresentado na Tabela 1. Foram obtidas as regiões de confiança utilizando-se “bootstrap” e “bootstrap” Bayesiano aliados a métodos não paramétricos de estimação de funções densidade de probabilidade e também por meio do cálculo da distância de Mahalanobis entre as coordenadas do ponto estacionário da amostra observada e aquelas obtidas por meio das estimativas “bootstrap”. Determinaram-se, ainda, as regiões de confiança de acordo com a metodologia de Box e Hunter (1954).

2.3 Resultados

2.3.1 Simulação

As porcentagens de cobertura com 90% e 95% de confiança para as coordenadas do ponto estacionário verdadeiro obtidas com os diferentes métodos podem ser observadas nas Tabelas 2, 3 e 4.

Na Tabela 2 são apresentados os valores referentes à utilização da metodologia de Box e Hunter (1954). Se forem considerados os quantis de ordem 2,5% e 97,5% da distribuição binomial com parâmetros 0,90 e 0,95, obtém-se, respectivamente, os intervalos de confiança (0,881; 0,918) e (0,936; 0,963). Nota-se que em todas as situações estudadas os intervalos contém as porcentagens de cobertura reais. Esta situação não era esperada no caso dos conjuntos de dados gerados com erros com distribuição lognormal, e demonstra que a metodologia clássica de Box e Hunter, baseada na pressuposição de que os erros têm distribuição normal é bastante robusta, mesmo em se tratando de situações em que a distribuição dos erros é bastante assimétrica.

Os resultados referentes à utilização dos métodos “bootstrap” aliada ao cálculo da distância de Mahalanobis entre as coordenadas do ponto estacionário da amostra observada e aquelas obtidas através das estimativas “bootstrap” podem ser visualizados na Tabela 3. Similarmente ao comportamento observado no método anterior, em todas as situações os intervalos de confiança contém as porcentagens de cobertura reais. Embora fosse esperado um melhor comportamento do método “bootstrap” Bayesiano em relação ao “bootstrap” clássico por, teoricamente, este último propiciar uma melhor cobertura no espaço paramétrico isto não ocorreu. Os resultados alcançados neste trabalho, discordam em parte daqueles obtidos por Freitas (1993), que em um estudo similar com o método “bootstrap” clássico verificou que a concordância entre os níveis nominais e reais ocorria apenas quando a superfície era bem definida (valores absolutos dos autovalores maiores do que o desvio padrão dos erros do modelo). Numa situação de uma superfície plana (valores absolutos dos autovalores menores do que o desvio padrão dos erros do modelo) o nível de significância real tendia a ser menor que o nominal, em torno de 5%, o que gera regiões de confiança mais amplas, com níveis

de confiança reais superiores aos nominais. No entanto salientou que o procedimento adotado gera regiões de confiança para a localização exata do ponto estacionário com razoável precisão.

Tabela 2 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de $\mathbf{B}$ e com a distribuição dos erros - Metodologia de Box e Hunter (1954)

λ_1	λ_2	Coberturas nominais	
		90%	95%
distribuição lognormal(0;1)			
-0,025	-0,017	89,3	93,7
-0,250	-0,170	90,0	95,1
-0,640	-0,590	90,3	95,0
-0,125	-0,117	89,1	94,1
distribuição normal(0;0,5)			
-0,025	-0,017	90,8	95,3
-0,250	-0,170	90,6	95,4
-0,640	-0,590	91,0	95,5
-0,125	-0,117	91,6	95,9
distribuição normal(0;0,8)			
-0,025	-0,017	89,7	94,1
-0,250	-0,170	89,5	94,1
-0,640	-0,590	90,6	94,6
-0,125	-0,117	89,9	95,9
distribuição normal(0;1)			
-0,025	-0,017	89,9	95,7
-0,250	-0,170	90,3	95,8
-0,640	-0,590	91,5	95,7
-0,125	-0,117	89,4	95,8

Os resultados referentes à utilização dos métodos “bootstrap” aliada a métodos

não paramétricos de estimação de funções densidade de probabilidade podem ser visualizados na Tabela 4.

Tabela 3 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de $\mathbf{B}$ e com a distribuição dos erros - Distância de Mahalanobis

λ_1	λ_2	Coberturas nominais			
		90%		95%	
		$BB^{(1)}$	$BC^{(2)}$	$BB^{(1)}$	$BC^{(2)}$
distribuição lognormal(0;1)					
-0,025	-0,017	88,3	89,9	94,6	95,3
-0,250	-0,170	90,5	88,9	94,1	93,8
-0,640	-0,590	89,2	90,4	94,2	95,0
-0,125	-0,117	91,3	88,7	96,0	93,6
distribuição normal(0;0,5)					
-0,025	-0,017	88,5	89,1	93,6	94,7
-0,250	-0,170	88,9	88,8	93,9	95,7
-0,640	-0,590	92,8	91,2	96,1	96,2
-0,125	-0,117	91,7	90,6	96,2	95,2
distribuição normal(0;0,8)					
-0,025	-0,017	88,5	90,5	94,5	95,2
-0,250	-0,170	88,7	89,8	93,7	95,4
-0,640	-0,590	90,1	89,5	94,3	94,4
-0,125	-0,117	90,9	90,2	96,2	95,0
distribuição normal(0;1)					
-0,025	-0,017	89,4	89,2	94,4	95,1
-0,250	-0,170	87,8	88,9	93,0	94,5
-0,640	-0,590	89,4	89,4	94,0	94,2
-0,125	-0,117	91,5	90,3	95,4	95,1

⁽¹⁾ “Bootstrap” Bayesiano e ⁽²⁾ “Bootstrap” clássico

Tabela 4 - Porcentagens de vezes em que as regiões de 90% e 95% cobriram o ponto estacionário verdadeiro, de acordo com os autovalores de $\mathbf{B}$ e com a distribuição dos erros - contorno de densidade não paramétrica bivariada

λ_1	λ_2	Coberturas nominais			
		90%		95%	
		$BB^{(1)}$	$BC^{(2)}$	$BB^{(1)}$	$BC^{(2)}$
distribuição lognormal(0;1)					
-0,025	-0,017	89,1	98,6	94,6	99,2
-0,250	-0,170	90,6	98,8	94,2	99,5
-0,640	-0,590	86,6	98,9	91,9	99,6
-0,125	-0,117	82,7	98,1	87,9	99,0
distribuição normal(0;0,5)					
-0,025	-0,017	94,7	98,8	97,3	99,4
-0,250	-0,170	94,4	98,7	97,7	99,2
-0,640	-0,590	84,6	96,1	89,9	97,9
-0,125	-0,117	68,8	90,8	78,3	94,0
distribuição normal(0;0,8)					
-0,025	-0,017	94,7	98,7	96,9	99,5
-0,250	-0,170	93,1	98,6	96,5	99,1
-0,640	-0,590	76,0	92,7	85,1	96,2
-0,125	-0,117	59,3	89,3	70,4	92,6
distribuição normal(0;1)					
-0,025	-0,017	94,7	99,4	97,6	99,6
-0,250	-0,170	93,6	98,6	96,2	99,3
-0,640	-0,590	90,6	97,7	94,4	98,2
-0,125	-0,117	72,0	91,3	80,2	95,2

⁽¹⁾ “Bootstrap” Bayesiano e ⁽²⁾ “Bootstrap” clássico

De acordo com os resultados observa-se que, exceto no caso do “bootstrap” Bayesiano sob distribuição lognormal LN(0,1) nas demais situações estudadas, para os autovalores com menor valor absoluto, o nível de significância real é menor do que o nominal o que origina regiões de confiança mais amplas. Este comportamento é mais freqüente em se tratando do “bootstrap” clássico. Por sua vez, à medida que aumenta o valor absoluto dos autovalores ocorre um cenário diferente, passando-se a ter regiões de confiança menores que a nominal, principalmente no caso do “bootstrap” Bayesiano. Apenas em poucas situações houve concordância entre os níveis de significância reais e nominais. Provavelmente isto deve-se ao modo de selecionar-se a constante de suavização, a qual, conforme Silverman (1986) e Bowman e Azalini (1997) afeta a contorno da densidade não paramétrica. Resultados similares foram observados por Silva e Zocchi (2004) para o “bootstrap” clássico que no entanto, usaram um procedimento diferente do adotado neste trabalho para a seleção da constante de suavização.

2.3.2 Análise dos dados de produção de amendoim

Para o ajuste do modelo de segundo grau os valores de $\mathbf{x}_1$ e $\mathbf{x}_2$ foram codificados por: $x_{1i} = (a_{1i} - 85)/35$ e $x_{2i} = (a_{2i} - 20)/5$ ($i = 1, \dots, 18$), produzindo as seguintes estimativas pontuais pelo método de mínimos quadrados: $\hat{\beta}_0 = 16,365$, $\hat{\mathbf{b}} = [1,6751 \ 2,7650]^T$,

$$\hat{\mathbf{B}} = \begin{bmatrix} -2,4634 & -0,1669 \\ -0,1669 & -1,9309 \end{bmatrix}$$

e $\hat{\boldsymbol{\theta}} = [0,2932 \ 0,6906]^T$, o que corresponde às quantidades de fertilizantes $a_1 = 95,26$ e $a_2 = 23,45$ libras por parcela. Os autovalores de $\hat{\mathbf{B}}$, $\lambda_1 = -1,8830$ e $\lambda_2 = -2,5114$ são menores do que zero e, portanto, $\hat{\boldsymbol{\theta}}$ é um ponto de máximo.

As estimativas de $\boldsymbol{\theta}$ e os respectivos erros padrões obtidos por meio dos métodos “bootstrap” e “bootstrap” Bayesiano são apresentados na Tabela 5.

As regiões de confiança de Box e Hunter com 90% e 95% de $\boldsymbol{\theta}$ podem ser observadas na Figura 1. As distribuições “bootstrap” Bayesiana de $\boldsymbol{\theta}$ e “bootstrap” de $\hat{\boldsymbol{\theta}}$ com as respectivas regiões de confiança de 90% e de 95% obtidas por meio da densidade não paramétrica e da distância de Mahalanobis podem ser visualizadas nas Figuras 2 e 3.

Tabela 5 - Estimativas de θ e erros padrões obtidos através dos métodos de mínimos quadrados, “bootstrap” clássico e “bootstrap” Bayesiano

Método	$\hat{\theta}_1$	$\hat{\theta}_2$
Mínimos quadrados	0,2932 (0,0855)	0,6902 (0,1824)
“bootstrap” clássico	0,2944 (0,0982)	0,7464 (0,2332)
“bootstrap” Bayesiano	0,2940 (0,0628)	0,7060 (0,0979)

Figura 1 - Regiões de 90% e de 95% de confiança para a localização do ponto estacionário utilizando o método de Box e Hunter (1954)

A partir da análise de variância para o modelo adotado, obteve-se $\hat{\sigma}^2 = 0,8688$ com $\nu = 12$ graus de liberdade.

Os resultados dos dois métodos baseados nas técnicas “bootstrap” são similares, o que é esperado, uma vez que se utilizou uma distribuição a priori não informativa, no caso do “bootstrap” Bayesiano. Pode-se verificar que as regiões de confiança determinadas pelos três métodos analisados estão localizadas dentro da região experimental. No entanto, observa-se que os valores das estimativas do “bootstrap” Bayesiano estão um pouco mais próximas das estimativas pontuais de mínimos quadrados, além de apresentarem menor dispersão; no caso de $\hat{\theta}_2$ o erro padrão obtido com o método “bootstrap” foi bem superior àquele obtido com o método “bootstrap” Bayesiano; esses fatos propiciaram a maior concentração de pontos em torno da estimativa de $\hat{\theta}_2$ apresentada pelo segundo método, diminuindo a área da região de

Figura 2 - Regiões de 90% e de 95% de confiança para a localização do ponto estacionário utilizando o método “bootstrap” com contornos da densidade não paramétrica bivariada - $\mathbf{h} = [0,0311 \ 0,0750]^T$ (esquerda) e da distância de Mahalanobis (direita) considerando-se $M = 1000$ amostras

Figura 3 - Regiões de 90% e de 95% de confiança para a localização do ponto estacionário utilizando o método “bootstrap” Bayesiano com contornos da densidade não paramétrica bivariada - $\mathbf{h} = [0,0199 \ 0,0310]^T$ (esquerda) e da distância de Mahalanobis (direita) considerando-se $M = 1000$ amostras

confiança, conforme se constata na análise dos gráficos apresentados nas Figuras 2 e 3.

Finalmente, é importante observar que, embora, os métodos “bootstrap” Bayesiano e “bootstrap” clássico dêem respostas similares, a interpretação destas respostas é bastante diferente. Enquanto o primeiro mostra a incerteza a posteriori sobre os parâmetros

de acordo com um modelo específico (multinomial/Dirichlet), o segundo estima a incerteza das estimativas de mínimos quadrados dos parâmetros (RUBIN,1981; LANCASTER, 2003).

3 CONSIDERAÇÕES FINAIS

Este trabalho abordou a construção de regiões de confiança para as coordenadas do ponto estacionário de uma superfície de resposta. Foram estudadas diferentes situações considerando-se a forma das superfícies analisadas, maior ou menor grau de curvatura, além de diferentes distribuições e magnitudes da variância dos erros. As regiões de confiança foram construídas por meio dos procedimentos de Box e Hunter (1954) (BH), “bootstrap” e “bootstrap” Bayesiano, aliado ao cálculo da distância de Mahalanobis entre as coordenadas do ponto estacionário da amostra observada e aquelas obtidas através das estimativas “bootstrap” (BM e BBM), e ainda, métodos “bootstrap” e “bootstrap” Bayesiano aliados a métodos não paramétricos de estimação de funções densidade de probabilidade (BNP e BBNP). A avaliação da metodologia foi realizada por meio de simulação e foi aplicada a um conjunto de dados de produção de amendoim.

No estudo de simulação a metodologia BH, baseada na distribuição normal dos erros, apresentou um bom desempenho em todas as situações analisadas, havendo concordância entre as regiões de confiança nominais e reais, mesmo naquelas em que essa distribuição é bastante assimétrica. Este mesmo comportamento foi observado para os métodos BM e BBM. No entanto, os métodos BNP e BBNP não apresentaram um desempenho satisfatório, resultando em um nível de significância real menor que o nominal para os autovalores com menor valor absoluto, gerando regiões de confiança maiores. No caso de autovalores com menor valor absoluto observou-se situação inversa. Baseado nestes resultados, tanto o método BH quanto os métodos BM e BBM são indicados para a construção das regiões de confiança para as coordenadas do ponto estacionário.

No caso da análise do conjunto de dados de amendoim os métodos BH, BM e BNP apresentaram regiões de confiança mais amplas comparativamente aos métodos BBM e BBNP. No entanto, os valores das estimativas do “bootstrap” Bayesiano são mais próximas das estimativas de mínimos quadrados e apresentam menor dispersão o que explica a menor área da região de confiança.

A metodologia proposta baseada no método “bootstrap” Bayesiano descrito por Rubin (1981), prescinde de qualquer conhecimento a priori. Considerando este fato é

interessante que se estude formas de incorporar alguma informação a priori na análise, o que possibilitaria o seu uso em inferências mais precisas. Também é importante que haja mais pesquisas sobre a forma de seleção da constante de suavização no cálculo de densidades não paramétricas multivariadas.

REFERÊNCIAS

- ANDRADE, D.F. **Alguns aspectos da metodologia de superfície de resposta**. 1977. 87 p. Dissertação (Mestrado em Estatística) - Instituto de matemática e estatística, Universidade de São Paulo, São Paulo, 1977.
- BÖCKENHOLT, U. Analyzing optima in the exploration of multiple response surfaces. **Biometrics**, Washington, v.45, p. 1001 – 1008, 1989
- BOWMAN, A.W.; AZZALINI, A. **Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustration**. New York: Oxford Universidade Press, 1997. 210 p.
- BOX, G.E.P; DRAPER, N.R. A basis for the selection of response surface design. **Journal of the American Statistical Association**, Washington, v.54, p. 622 – 654, 1959
- BOX, G.E.P; DRAPER, N.R. The choice of a second order rotatable design. **Biometrika**, London, v.50, p. 335 – 352, 1963
- BOX, G.E.P; DRAPER, N.R. **Empirical model building and response surfaces**. New York: John Wiley, 1987. 669 p.
- BOX, G.E.P; HUNTER, J.S. A confidence region for the solution of a set de simultaneous equations with an application to experimental design. **Biometrika**, Oxford, v.40, p. 190 – 199, 1954
- BOX, G.E.P; HUNTER, J.S. Multifactor experimental designs for exploring response surfaces. **The Annals of Mathematical Statistics**, Stanford, v.28, p. 195 – 241, 1957
- BOX, G.E.P; LUCAS, H.L. Design of experiments in nonlinear situations. **Biometrika**, London, v.46, p. 77 – 90, 1959
- BOX, G.E.P; WILSON, K.B. 1951 On the experimental attainment of optimum conditions (with discussion). **Journal of the Royal Statistical Society. Series B**, Oxford, v.13, p. 1 – 45, 1951
- CAHYA, S.; CASTILLO, E. del; PETERSON, J.J. Computation of confidence regions for optimal factor levels in constrained response surface problems. **Journal of computational and graphical statistics**, New York, v.13, p. 499 – 518, 2004
- CAMPOS, H. de **Aspectos da aplicação das superfícies de resposta a ensaios fatoriais 3^3 de adubação**. 1967. 82 p. Tese (Livre-Docência) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1967.
- CARTER, W.H.; CHINCHILLI, V.M.; CAMPBELL, E.D.; WAMPLER, G.L. Confidence interval about the response at the stationary point of a response surface, wity an application to preclinical cancer thrapy. **Biometrics**, Washington, v.40, p. 1135 – 1130, 1984
- CASTILLO, E. del; CAHYA, S. A tool for computing confidence regions on the stationary point of a response surface. **The american statistician**, Washington, v.55, p. 358 – 365, 2001
- DAVISON, A.C.; HINKLEY, D.V. **Bootstrab methods and their application**. Cambridge: Cambridge University Press, 1997. 582p

EFRON, B. Bootstrap methods: Another look at the jackknife. **Annals of Statistics**, Washington, v.7, p.1 – 26, 1979.

EFRON, B. **The Jackknife, the Bootstrap and other Resampling Plans**. Philadelphia: SIAM, 1982. v.38, 93p

EFRON, B.; TIBSHIRANI, R.J. **An introduction to the Bootstrap**. London: Chapman & Hall, 1993. 436p.

FERGUSON, T.S. A bayesian analysis of some nonparametric problems. **Annals of Statistics**, Hayward, v.1, p.209 – 230, 1973.

FREITAS, S.M. **DE Emprego do bootstrap para determinação de regiões de confiança para os localizadores do ponto ótimo de operação em modelos de superfície de respostas**. 1993. 81 p. Dissertação (Mestrado em Estatística) - Instituto de matemática, estatística e ciência da computação, Universidade de Campinas, Campinas, 1993.

GILMOUR, S.G.; DRAPER, N.R. Confidence regions around the ridge of optimal response on fitted second order response surfaces. **Technometrics**, Alexandria, v.45, p.333 – 339, 2003.

GOMES, M. B. **Intervalos de confiança para as coordenadas do ponto crítico da equação quadrática de regressão a duas variáveis independentes**. 1989. 70 p. Tese (Livre-Docência) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1989.

KHURI, A.I.; CORNELL, J.A. **Response Surfaces - Designs and analysis**. 2.ed. New York: Marcel Dekker, 1996. 510 p.

LANCASTER, T. **A note on bootstraps and robustness**. Disponível em: < [http : //www.econ.brown.edu/ ~ tl/papers/robust.pdf](http://www.econ.brown.edu/~tl/papers/robust.pdf) >. Acesso em: 10 dez. 2007.

MALHEIROS, E. B.; PERECIN, D. Posição das doses de nutrientes na análise de uma superfície de resposta à adubação. **Revista de Matemática e Estatística**, Jaboticabal, v.15, p. 69 – 78, 1983

MANLY, B.F.J. **Randomization, Bootstrap and Monte Carlo Methods in Biology**. 2ª ed. London: Chapman & Hall, 1997. 399p.

MEAD, R.; PIKE, D.J. A review of response surface methodology from a biometric viewpoint. **Biometrics**, Washington, v.31, p. 803 – 851, 1975

MYERS, R.H.; KHURI, A.L.; CARTER Jr., W.H. Response surface methodology: 1966-1988. **Technometrics**, Alexandria, v.31, p. 137 – 157, 1989

MYERS, R.H. Response surface methodology: current status and future directions. **Journal of Quality Technology**, Milwaukee, v.31, p. 30 – 44, 1999

MYERS, R.H.; MONTGOMERY, D.C. **Response Surface Methodology**. 2.ed. New York: John Wiley, 2002, 798 p.

MYERS, R.H.; MONTGOMERY, D.C.; VINING, G.G.; BORROR, C.M.; KOWALSKI, S.M. Response surface methodology: a retrospective and literature survey. **Journal of Quality Technology**, Milwaukee, v.36, p. 53 – 77, 2004

PETERSON, J.J.; CAHYA, S.; CASTILLO, E. del. A general approach to confidence regions for optimal factor levels of response surfaces. **Biometrics**, Washington, v.58, p. 422 – 431, 2002

PIMENTEL GOMES, F. **Curso de estatística experimental**. Piracicaba: Nobel, 1990, 468 p.

PIMENTEL GOMES, F.; CONAGIN, A. Experimentos de adubação: planejamento e análise estatística. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA-REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 1987, Londrina. **Anais...** Londrina: UEL, 1987. Curso, 102 p.

R Development Core Team. **R: A language and environment for statistical computing**. Vienna, 2006. URL. Disponível em: < <http://www.R-project.org> >. Acesso em: 17 out. 2006.

RUBIN, D.B. Bayesian bootstrap. **Annals of Statistics**, Washington, v.9, p.130 – 134, 1981.

SANCHES, A. **Superfícies de resposta em experimentos de adubação: o problema dos pontos de sela**. 1986. 91 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1986.

SILVA, R.R.; ZOCCHI, S.S. Regiões de confiança para a localização do ponto estacionário em superfícies de resposta usando “bootstrap” e métodos “kernel”, com escolha automática das constantes de suavização. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 49., 2004, Uberlândia. **Resumos...** Uberlândia: UFU, 2004. Resumo 108.

SILVERMAN, B.W. **Density estimation for statistics and data analysis**. London: Chapman & Hall, 1986, 175 p.

STABLEIN, D.M.; CARTER, W.H.; WAMPLER, G.L. Confidence regions for constrained optima in response surface experiments. **Biometrics**, Washington, v.39, p. 754 – 763, 1983

WILKS, S.S. **Mathematical statistics**. New York: John Wiley, 1962, 644p.

ZOCCHI, S.S.; SILVA, R.R. Regiões de confiança para a localização do ponto estacionário em superfícies de resposta usando bootstrap. In: SIMPÓSIO DE ESTATÍSTICA APLICADA À EXPERIMENTAÇÃO AGRONÔMICA, 10., 2003, Lavras. **Resumos...** Lavras: UFLA, 2003. Resumo 305.

APÊNDICES

APÊNDICE A - Programas utilizados na simulação e na análise dos dados

#SIMULAÇÃO MONTE CARLO - CONJUNTO DE DADOS

```
x1<-cbind(c(-1,1,-1,1,0,0,0,0,-sqrt(2),sqrt(2),0,0))
x2<-cbind(c(-1,-1,1,1,0,0,0,0,0,0,-sqrt(2),sqrt(2)))
X<-cbind(rep(1,times=12),x1,x2,x1^2,x2^2,x1*x2)
M<-cbind(c(cos(pi/6),sin(pi/6)),c(-sin(pi/6),cos(pi/6)))
Lambda<-cbind(c(-1.25,0),c(0,-1.17))
Theta<-cbind(rep(0.2,times=12),rep(0.4,times=12))
yobs<-function(X,M,Lambda,Theta,RES){
y1<-(X[,2:3]-Theta)%*M%*Lambda%*t(M)%*t(X[,2:3]-Theta)+
rep(100,times=12) y2<- cbind(diag(y1)) y<- RES+y2 y }
y<-rbind(numeric(1000),numeric(1000),numeric(1000),numeric(1000),
numeric(1000),numeric(1000),numeric(1000),numeric(1000),
numeric(1000),numeric(1000),numeric(1000),numeric(1000)) \\
for(i in 1:1000)
y[,i]<-yobs(X,M,Lambda,Theta,cbind(rnorm(12,0,1)))
save(y,file="D:\\ARQ_SIMUL\\y_norm1_l125.Rdata")
```

#GERA AMOSTRA BOOSTRAP BAYESIANO

```
x1<-cbind(c(-1,1,-1,1,0,0,0,0,-sqrt(2),sqrt(2),0,0))
x2<-cbind(c(-1,-1,1,1,0,0,0,0,0,0,-sqrt(2),sqrt(2)))
X<-cbind(rep(1,times=12),x1,x2,x1^2,x2^2,x1*x2)
load("D:\\ARQ_SIMUL\\y_norm1_l125.Rdata")
beta<-solve(crossprod(X),crossprod(X,y))
```

Bootstrap

```
mmq<-function(X,w){ a1<-X a2<-w mq<-function(a1,a2){
```



```

un<-sort(runif(11)) prob<-c(un,1)-c(0,un) g<-diag(prob)
B1<-solve(t(a1)%*%g%*%a1)%*%t(a1)%*%g%*%a2
B1
est<-solve(cbind(c(B1[4,1],B1[6,1]/2),c(B1[6,1]/2,B1[5,1])))%*%
cbind(c(B1[2,1],B1[3,1]))*-0.5 est }
estac<-rbind(numeric(1000),numeric(1000)) for(i in 1:1000)
param[,i]<- mq(a1,a2) estac } estac<-matrix(c(2000000),2000,1000)
for(i in 1:1000) estac[,i]<-mmq(X,y[,i])
x2otimo<-matrix(c(numeric(1000000)),1000,1000) for(i in 1:1000)
x2otimo[i,]<-estac[2*i,]
x1otimo<-matrix(c(numeric(1000000)),1000,1000) for(i in 1:1000)
x1otimo[i,]<-estac[2*i-1,]
save(x1otimo,x2otimo, file =
"D:\\ARQ_SIMUL\\y_norm1_l125_bayes.Rdata")

# GERA AMOSTRA BOOSTRAP CLÁSSICO

x1<-cbind(c(-1,1,-1,1,0,0,0,0,-sqrt(2),sqrt(2),0,0))
x2<-cbind(c(-1,-1,1,1,0,0,0,0,0,0,-sqrt(2),sqrt(2)))
X<-cbind(rep(1,times=12),x1,x2,x1^2,x2^2,x1*x2)
load("D:\\DAVID\\ARQ_SIMUL_2\\y_norm01_l125.Rdata")
beta<-solve(crossprod(X),crossprod(X,y))
res<-y-X%*%beta

#Bootstrap

mmq<-function(X,w,resid){ a1<-X a2<-w a3<-resid
mq<-function(a1,a2,a3){
rcorr<-a3%*%matrix(c(sqrt(12/(12-6))),1,1)
yobs<-a1%*%a2 +rcorr

```

```

B1<-solve(crossprod(a1),crossprod(a1,yobs)) B1
est<-solve(cbind(c(B1[4,1],B1[6,1]/2),c(B1[6,1]/2,B1[5,1])))%*%
cbind(c(B1[2,1],B1[3,1]))*-0.5 est }
estac<-rbind(numeric(1000),numeric(1000)) for(i in 1:1000)
param[,i]<- mq(a1,a2,sample(a3,replace=TRUE)) }
estac<-matrix(c(2000000),2000,1000) for(i in 1:1000)
estac[,i]<-mmq(X,beta[,i],res[,i])
x2otimo<-matrix(c(numeric(1000000)),1000,1000) for(i in 1:1000)
x2otimo[i,]<-estac[2*i,]
x1otimo<-matrix(c(numeric(1000000)),1000,1000) for(i in 1:1000)
x1otimo[i,]<-estac[2*i-1,]
save(x1otimo,x2otimo, file =
"D:\\ARQ_SIMUL\\y_norm1_l125_efron.Rdata")

```

CALCULO DA PORCENTAGEM DE COBERTURA BOX-HUNTER

```

x1<-cbind(c(-1,1,-1,1,0,0,0,0,-sqrt(2),sqrt(2),0,0))
x2<-cbind(c(-1,-1,1,1,0,0,0,0,0,0,-sqrt(2),sqrt(2)))
X<-cbind(rep(1,times=12),x1,x2,x1^2,x2^2,x1*x2) theta1<-c(0.2,0.4)
load("D:\\ARQ_SIMUL\\y_norm1_l125.Rdata")
XTX<-solve(t(X)%*%X)
xty<-t(X)%*%y
beta<-XTX%*%xty
vdel1<-XTX[2,2]+4*XTX[4,4]*theta1[1]^2+XTX[6,6]*theta1[2]^2
vdel2<-XTX[3,3]+XTX[6,6]*theta1[1]^2+4*XTX[5,5]*theta1[2]^2
covdel12<-4*XTX[4,5]*theta1[1]*theta1[2]+XTX[6,6]*theta1[1]*theta1[2]
v<-matrix(c(vdel1,covdel12,covdel12,vdel2),2,2,byrow=T)
delta1<-NULL delta2<-NULL for(i in 1:1000) {
delta1[i]<-beta[2,i]+2*beta[4,i]*theta1[1]+beta[6,i]*theta1[2]
delta2[i]<-beta[3,i]+2*beta[5,i]*theta1[2]+beta[6,i]*theta1[1]}

```

```

delta<-matrix(rbind(delta1,delta2),2,1000)
regconf<-t(delta)%*%solve(v)%*%delta
regconf<-diag(regconf)
res<-matrix(c(y-X%*%beta),12,1000)
mse<-NULL
mse<-1/6*diag(t(res)%*%res)
lconf90<-qf(0.90,2,6)*2*mse
lconf95<-qf(0.95,2,6)*2*mse
cob90<-regconf<=lconf90
cob95<-regconf<=lconf95
cob_90<-sum(cob90)
cob_95<-sum(cob95)
cob_90
cob_95

```

CÁLCULO DA PORCENTAGEM DE COBERTURA DENSIDADE NÃO PARAMÉTRICA

```

load("D:\\ARQ_SIMUL\\y_norm1_l125_efron.Rdata") y1<-x1otimo
y2<-x2otimo est1<-matrix(rep(0.2,times=1000),1000,1000)
est2<-matrix(rep(0.4,times=1000),1000,1000) library(sm)
dens<-function(x,y,x1,y1){ h1<-(1/1000)^(1/6)*sd(x)
h2<-(1/1000)^(1/6)*sd(y)
d<-sm.density(cbind(x,y),h=c(h1,h2),eval.grid=T,
eval.points=cbind(x1,y1),display="none")
quantile(d$estimate,probs=c(0.10)) } r1<-numeric(1000) for( i in
1:1000)
r1[i]<-dens(cbind(y1[,i]),cbind(y2[,i]),cbind(y1[,i]),cbind(y2[,i]))
r2<-numeric(1000) for( i in 1:1000)
r2[i]<-dens(cbind(y1[,i]),cbind(y2[,i]),cbind(est1[,i]),cbind(est2[,i]))
resul<-r2>r1 cob90<-sum(resul) dens<-function(x,y,x1,y1){

```

```

h1<-(1/1000)^(1/6)*sd(x) h2<-(1/1000)^(1/6)*sd(y)
d<-sm.density(cbind(x,y),h=c(h1,h2),eval.grid=T,
eval.points=cbind(x1,y1),display="none")
quantile(d$estimate,probs=c(0.05)) } r1<-numeric(1000) for( i in
1:1000)
r1[i]<-dens(cbind(y1[,i]),cbind(y2[,i]),cbind(y1[,i]),cbind(y2[,i]))
r2<-numeric(1000) for( i in 1:1000)
r2[i]<-dens(cbind(y1[,i]),cbind(y2[,i]),cbind(est1[,i]),cbind(est2[,i]))
resul<-r2>=r1 cob95<-sum(resul)
cob90
cob95

```

CALCULO DA PORCENTAGEM DE COBERTURA DISTÂNCIA DE MAHALANOBIS

```

load("D:\\DAVID\\ARQ_SIMUL_2\\y_norm08_l125_efron.Rdata")
y1<-x1otimo y2<-x2otimo
est1<-matrix(rep(0.2,times=1000),1000,1000)
est2<-matrix(rep(0.4,times=1000),1000,1000)
distmah<-function(x1,x2,x1e,x2e){ xm<-matrix(c(x1,x2),1000,2)
cvar<-cov(xm)
dist<-matrix(c(rep(x1e-mean(x1),times=1000),rep(x2e-mean(x2),times=1000)),
1000,2)
distmal<-dist%*%solve(cvar)%*%t(dist)
dist1<-diag(distmal) quantile(dist1,prob=0.10) } r1<-numeric(1000)
for( i in 1:1000)
r1[i]<-distmah(cbind(y1[,i]),cbind(y2[,i]),cbind(y1[,i]),cbind(y2[,i]))
r2<-numeric(1000) for( i in 1:1000)
r2[i]<-distmah(cbind(y1[,i]),cbind(y2[,i]),cbind(est1[,i]),cbind(est2[,i]))
resul<-r2>=r1 cob90<-sum(resul)
distmah<-function(x1,x2,x1e,x2e){

```

```

xm<-matrix(c(x1,x2),1000,2) cvar<-cov(xm)
dist<-matrix(c(rep(x1e-mean(x1),times=1000),rep(x2e-mean(x2),times=1000)),
1000,2)
distmal<-dist%*%solve(cvar)%*%t(dist)
dist1<-diag(distmal) quantile(dist1,prob=0.05) } r1<-numeric(1000)
for( i in 1:1000)
r1[i]<-distmah(cbind(y1[,i]),cbind(y2[,i]),cbind(y1[,i]),cbind(y2[,i]))
r2<-numeric(1000) for( i in 1:1000)
r2[i]<-distmah(cbind(y1[,i]),cbind(y2[,i]),cbind(est1[,i]),cbind(est2[,i]))
resul<-r2>=r1 cob95<-sum(resul)
cob90
cob95

```

GERA AMOSTRAS BOOTSTRAP BAYES E EFRON - AMENDOIM

```

xdel<-matrix(c(-1,-1,1,1,-1,-1,1,1,-sqrt(2),-sqrt(2),sqrt(2),sqrt(2),0,0,
0,0,0,0,-1,-1,-1,-1,1,1,1,1,0,0,0,0,-sqrt(2),-sqrt(2),sqrt(2),sqrt(2),0,0)
,18,2)
y<-matrix(c(7.52,8.12,12.37,11.84,13.55,12.35,16.48,15.32,8.63,9.44,14.22,
12.57,7.90,7.33,16.49,17.40,15.73,17),18,1)
X<-cbind(rep(1,18),xdel,xdel^2,xdel[,1]*xdel[,2])
beta<-solve(crossprod(X),crossprod(X,y)) b<-c(beta[2],beta[3])
B<-matrix(c(beta[4],beta[6]/2,beta[6]/2,beta[5]),2,2)
estac<--1/2*solve(B)%*%b
const<-matrix(cbind(rep(estac[1,1],times=1000),rep(estac[2,1],times=1000))
,1000,2)

#EFRON

res<-y-X%*%beta

```

```

rescor<-res*sqrt(2)
estac.efron<-matrix(NA,1000,2) for (i in
1:1000) { amos.res<-sample(rescor,replace=TRUE) ye2<-y+amos.res
b2<-solve(crossprod(X),crossprod(X,ye2)) bhat<-c(b2[2],b2[3])
Bhat<-matrix(c(b2[4],b2[6]/2,b2[6]/2,b2[5]),2,2)
estac.efron[i,]<--1/2*solve(Bhat)%*%bhat}
save(estac.efron, file =
"D:\\ARQ_SIMUL\\efron_dados_amendoim.Rdata")

```

#BAYES

```

estac.rubin<-matrix(NA,1000,2) for (i in 1:1000){
un<-sort(runif(17)) prob<-c(un,1)-c(0,un) g<-diag(prob)
xxe2<-t(X)%*%g%*%X
xxye2<-t(X)%*%g%*%y
xxeinv2<-solve(xxe2)
b3<-xxeinv2%*%xxye2
bhat2<-c(b3[2],b3[3])
Bhat2<-matrix(c(b3[4],b3[6]/2,b3[6]/2,b3[5]),2,2)
estac.rubin[i,]<--1/2*solve(Bhat2)%*%bhat2
} save(estac.rubin, file =
"D:\\ARQ_SIMUL\\bayes_dados_amendoim.Rdata")

```

GRÁFICO REGIAO BH - AMENDOIM

```

x<-matrix(c(-1,-1,1,1,-1,-1,1,1,-sqrt(2),-sqrt(2),sqrt(2),sqrt(2),0,0,
0,0,0,0,-1,-1,-1,-1,1,1,1,1,0,0,0,0,-sqrt(2),-sqrt(2),sqrt(2),sqrt(2),
0,0),18,2) X<-cbind(rep(1,times=18),x,x^2,x[,1]*x[,2])
y<-c(7.52,8.12,12.37,11.84,13.55,12.35,16.48,15.32,8.63,9.44,14.22,
12.57,7.9,7.33,16.49,17.40,15.73,17)

```

```

XTX<-solve(crossprod(X))
xty<-crossprod(X,y)
beta<-XTX%%xty
theta1<-seq(0,1,length=100) theta2<-seq(0,2,length=100)
Theta<-expand.grid(theta1,theta2)
vdel1<-XTX[2,2]+4*XTX[4,4]*Theta[,1]^2+XTX[6,6]*Theta[,2]^2
vdel2<-XTX[3,3]+XTX[6,6]*Theta[,1]^2+4*XTX[5,5]*Theta[,2]^2
covdel12<-4*XTX[4,5]*Theta[,1]*Theta[,2]+XTX[6,6]*Theta[,1]*Theta[,2]
delta1<-beta[2]+2*beta[4]*Theta[,1]+beta[6]*Theta[,2]
delta2<-beta[3]+2*beta[5]*Theta[,2]+beta[6]*Theta[,1]
delta<-matrix(rbind(delta1,delta2),2,10000)
regcon<-matrix(0,100,100) for (i in 1:10000) {
v<-matrix(c(vdel1[i],covdel12[i],covdel12[i],vdel2[i]),2,2,byrow=T)
regcon[i]<-t(delta[,i])%%solve(v)%%delta[,i]}
res<-matrix(c(y-X%%beta),18,1)
mse<-1/12*diag(crossprod(res)) lconf90<-qf(0.90,2,12)*2*mse
lconf95<-qf(0.95,2,12)*2*mse
plot(theta1,theta2,type="n",bty="l",xlab=expression(hat(theta[1])),
ylab=expression(hat(theta[2])),las=1)
contour(theta1,theta2,regcon,levels=c(lconf90,lconf95),
labels=c("90","95"),lwd=1.5,add=T) points(0.293214,0.690621,pch=3)

```

GRÁFICO REGIAO MAHALANOBIS - AMENDOIM

#EFRON

```

xdel<-matrix(c(-1,-1,1,1,-1,-1,1,1,-sqrt(2),-sqrt(2),sqrt(2),sqrt(2),
0,0,0,0,0,0,-1,-1,-1,-1,1,1,1,1,0,0,0,0,-sqrt(2),-sqrt(2),sqrt(2),
sqrt(2),0,0),18,2)
y<-matrix(c(7.52,8.12,12.37,11.84,13.55,12.35,16.48,15.32,8.63,9.44,

```

```

14.22,12.57,7.90,7.33,16.49,17.40,15.73,17),18,1)
X<-cbind(rep(1,18),xdel,xdel^2,xdel[,1]*xdel[,2])
beta<-solve(crossprod(X),crossprod(X,y)) b<-c(beta[2],beta[3])
B<-matrix(c(beta[4],beta[6]/2,beta[6]/2,beta[5]),2,2)
estac<--1/2*solve(B)%*%b
const<-matrix(cbind(rep(estac[1,1],times=1000),rep(estac[2,1],
times=1000)),1000,2)
load("D:\\ARQ_SIMUL\\efron_dados_amendoim.Rdata")
cvar.efron<-cov(estac.efron)
dist<-matrix(c(estac.efron-const),1000,2)
distmal<-dist%*%solve(cvar.efron)%*%t(dist)
dist1<-diag(distmal) d1.efron<-quantile(dist1,prob=0.90)
d2.efron<-quantile(dist1,prob=0.95)
plot(estac.efron[,1],estac.efron[,2],xlim=c(0,1),ylim=c(0,2),
pch=".",col="black",xlab=expression(hat(theta[1])),
ylab=expression(hat(theta[2])),bty="l",las=1)
points(x=estac[1,1],y=estac[2,1],pch=20,col="black")
theta1<-seq(0,1,length=50) theta2<-seq(0,2,length=50)
Theta1<-expand.grid(theta1,theta2) attach(Theta1)
Theta2<-matrix(cbind(Var1,Var2),2500,2)
const2<-matrix(cbind(rep(estac[1,1],times=2500),rep(estac[2,1],
times=2500)),2500,2) dist2<-matrix(c(Theta2-const2),2500,2)
distmal2<-dist2%*%solve(cvar.efron)%*%t(dist2)
dist3<-matrix(diag(distmal2),50,50)
contour(theta1,theta2,dist3,levels=c(d1.efron,d2.efron),
labels=c("90","95"),lwd=1.5,add=T)

#BAYES

load("D:\\ARQ_SIMUL\\bayes_dados_amendoim.Rdata")

```



```

cvar.rubin<-cov(estac.rubin)
dist.rubin<-matrix(c(estac.rubin-const),1000,2)
distmal.rubin<-dist.rubin%%solve(cvar.rubin)%*%t(dist.rubin)
dist1.rubin<-diag(distmal.rubin)
d1.rubin<-quantile(dist1.rubin,prob=0.90)
d2.rubin<-quantile(dist1.rubin,prob=0.95)
plot(estac.rubin[,1],estac.rubin[,2],xlim=c(0,1),ylim=c(0,2),
pch=".",col="black",xlab=expression(theta[1]),
ylab=expression(theta[2]),bty="l",las=1)
points(x=estac[,1],y=estac[,2],pch=20,col="black")
distmal2<-dist2%%solve(cvar.rubin)%*%t(dist2)
dist3<-matrix(diag(distmal2),50,50)
contour(theta1,theta2,dist3,levels=c(d1.rubin,d2.rubin),
labels=c("90","95"),lwd=1.5,add=T)

```

GRÁFICO REGIAO DENSIDADE NÃO PARAMÉTRICA - AMENDOIM

```

load("D:\\ARQ_SIMUL\\bayes_dados_amendoim.Rdata")
Média.Rubin<-matrix(c(mean(estac.rubin[,1]),
mean(estac.rubin[,2])),2,1)
Desvio.padrão.Rubin<-matrix(c(sd(estac.rubin[,1]),
sd(estac.rubin[,2])),2,1)

load("D:\\ARQ_SIMUL_2\\efron_dados_amendoim.Rdata")
Média.Efron<-matrix(c(mean(estac.efron[,1]),
mean(estac.efron[,2])),2,1)
Desvio.padrão.Efron<-matrix(c(sd(estac.efron[,1]),
sd(estac.efron[,2])),2,1) Média.Efron Desvio.padrão.Efron
Média.Rubin Desvio.padrão. Rubin library(sm)
h1ef<-(1/1000)^(1/6)*sd(estac.efron[,1])

```

```
h2ef<-(1/1000)^(1/6)*sd(estac.efron[,2])
h1bay<-(1/1000)^(1/6)*sd(estac.rubin[,1])
h2bay<-(1/1000)^(1/6)*sd(estac.rubin[,2]) par(bty="l",las=1)
sm.density(estac.efron,display="slice",h=c(h1ef,h2ef),
xlab=expression(hat(theta[1])),ylab=expression(hat(theta[2])),
xlim=c(0,1),ylim=c(0,2),props=c(90,95))
sm.density(estac.rubin,display="slice",h=c(h1bay,h2bay),
xlab=expression(theta[1]),ylab=expression(theta[2]),
xlim=c(0,1),ylim=c(0,2),props=c(90,95))
```