

UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP

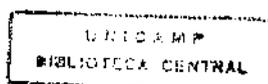
Instituto de Matemática, Estatística e Ciência da Computação - IMECC

EMPREGO DO *BOOTSTRAP* PARA DETERMINAÇÃO DE REGIÕES DE CONFIANÇA
PARA OS LOCALIZADORES DO PONTO ÓTIMO DE OPERAÇÃO EM MODELOS
DE SUPERFÍCIE DE RESPOSTAS

SÍLVIA MARIA DE FREITAS \checkmark 884
PROF. DR. SEBASTIÃO DE AMORIM \dagger
orientador

CAMPINAS - SÃO PAULO

1993



EMPREGO DO *BOOTSTRAP* PARA DETERMINAÇÃO DE REGIÕES DE CONFIANÇA
PARA OS LOCALIZADORES DO PONTO ÓTIMO DE OPERAÇÃO EM MODELOS
DE SUPERFÍCIE DE RESPOSTAS

Este exemplar corresponde à redação final da tese devidamente corrigida e defendida pela Srta. SILVIA MARIA DE FREITAS e aprovada pela Comissão Julgadora.

Campinas, 16 de julho de 1993.

Prof. Dr.



SEBASTIÃO DE AMORIM

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

dos meus pais,

Messias e Irene.

AGRADECIMENTOS

Agradeço à minha família pelo apoio, incentivo e compreensão dedicados à elaboração deste trabalho nos momentos de grandes preocupações.

Agradeço , especialmente, ao meu orientador Prof. Dr. Sebastião de Amorim pela orientação desempenhada para este trabalho e principalmente pelo apoio e incentivo demonstrado nos momentos de dificuldades.

Aos meus colegas antigos e às amizades adquiridas no Mestrado, agradeço pelo apoio e pelo incentivo em todos os momentos.

Ao Departamento de Estatística da Universidade Federal do Ceará, pela compreensão e apoio para a conclusão deste trabalho.

Agradeço aos membros da Banca Examinadora pela disponibilidade para a colaboração de participação na apresentação do trabalho.

Ao Departamento de Estatística da UNICAMP, em especial aos professores, pela colaboração na elaboração deste trabalho.

Ao Conselho Nacional de Desenvolvimento à Pesquisa, CNPq, pela participação a mim concedida em seus trabalhos de incentivo à pós-graduação em forma de bolsas de estudo.

RESUMO

As técnicas em Exploração de Superfícies de Respostas foram introduzidas pelo Prof Box no início da década de 50. Tais técnicas se constituem até hoje numa ferramenta bastante popularizada na abordagem de problemas de otimização de processos industriais. Em tais modelos, a busca pelo ponto ótimo de operação se conclui com a estimação das coordenadas do ponto ótimo. A dificuldade na análise desses modelos se encontra no fato da estimativa ser uma superfície. Neste trabalho exploramos a utilidade do método *Bootstrap*, introduzido por Efron em 1979, para permitir uma avaliação da variabilidade associada à estimativa dessa superfície.

No capítulo 1 é feita uma revisão geral na Metodologia de Superfície de Respostas, com a descrição da terminologia e os modelos utilizados, como também suas propriedades desejadas, utilizados para o ajuste e estimação da superfície.

No capítulo 2, a descrição do método *Bootstrap* é feita juntamente com suas aplicações em modelos de Regressão Linear e sua aplicabilidade no contexto de Superfície de Respostas.

No capítulo 3 são apresentados resultados obtidos por simulação, com delineamentos rotacionáveis de 2ª ordem para dois preditores. Em todos os casos explora-se a performance do *Bootstrap* no estabelecimento de regiões de confiança para as coordenadas do ponto ótimo de operação. É sempre adotado o modelo normal para os erros com diversos valores para σ .

Em anexo seguem os procedimentos de programação utilizados para as simulações desenvolvidas no decorrer do trabalho.

CAPÍTULO 2 - O "BOOTSTRAP" , UMA VISÃO GERAL E APLICAÇÕES NO CONTEXTO DE SUPERFÍCIE DE RESPOSTAS	35
2.1 O método "bootstrap"	35
2.2 O "bootstrap" em Regressão Linear	41
2.3 O "bootstrap" no contexto de Superfície de Respostas	45
CAPÍTULO 3 - AS SIMULAÇÕES EM MODELOS DE 2ª ORDEM	53
3.1 A simulação em modelos de 2ª ordem com dois preditores	53
3.2 Conclusões	70
ANEXO 1 - PROGRAMA	72
BIBLIOGRAFIA	79

CAPÍTULO 1 - A METODOLOGIA DE EXPLORAÇÃO DE SUPERFÍCIE DE RESPOSTAS

1.1 INTRODUÇÃO :

" Metodologia de Superfície de Respostas " (MSR) é um conjunto de técnicas matemáticas e estatísticas utilizadas para a modelagem e análise de problemas físicos, onde uma determinada variável resposta é influenciada de alguma forma por um conjunto de variáveis preditoras. Inicialmente introduzida por Box [1] e [2], em 1951 e 1954 respectivamente, teve nas décadas subsequentes uma atenção muito grande por parte não só de estatísticos e engenheiros, mas despertou muito interesse em pesquisadores de praticamente todas as áreas, incluindo-se a sociologia, a psicologia e a pedagogia.

Para um caso envolvendo k variáveis preditoras $\xi_1, \xi_2,$

..., ξ_k , a MSR assume a existência de uma relação determinística

$$\eta = f(\xi_1, \xi_2, \dots, \xi_k),$$

tal que a resposta y é dada por

$$y = \eta + \varepsilon,$$

onde ε é um desvio aleatório com uma distribuição com média 0 e desvio padrão $\sigma = \sigma(\xi_1, \xi_2, \dots, \xi_k)$. Em uma situação particular de interesse, $\varepsilon \sim N(0, \sigma^2)$.

Se, por um lado, a determinação da forma funcional f é de grande interesse no entendimento dos mecanismos internos do processo, por outro, num ambiente de produção, o interesse pode ser mais pragmático, concentrando-se não na função f , mas na determinação do particular valor de $(\xi_1, \xi_2, \dots, \xi_k)$ onde η é máximo (ou mínimo). Em uma palavra, o que se busca com a MSR é o ponto ótimo de operação.

A MSR busca trazer soluções para este segundo problema. Essa restrição de objetivos permite ao experimentador uma simplificação crucial. Sem a necessidade de uma modelagem global, em cada ponto de operação o que se deseja determinar é apenas um modelo empírico que se ajuste à superfície de respostas η em uma vizinhança mais ou menos estreita deste ponto. Esta simplificação é muito importante pois permite, por um lado, planos experimentais envolvendo perturbações mínimas nas variáveis preditoras, uma condição essencial em se tratando de experimentos em escala de produção; por outro lado,

modelando setorialmente apenas vizinhanças estreitas de um dado ponto central de operações, modelos empíricos polinômiais de baixa ordem - essencialmente de 1ª e 2ª ordens - são geralmente suficientes para oferecer descrições operacionais úteis da superfície de respostas na vizinhança considerada.

A MSR é um procedimento sequencial que se inicia com a modelagem por um modelo de primeira ordem da superfície de respostas na vizinhança do ponto usual de operações.

$$\eta = \alpha_0 + \sum_{i=1}^k \alpha_i \xi_i$$

onde os α_i 's são os coeficientes da i-ésima variável preditora. O ajuste é feito com base nas observações experimentais decorrentes da aplicação de um delineamento experimental de primeira ordem centrado no ponto de operações. Testada a adequação do modelo planar, determina-se a direção ótima de deslocamento do ponto de operação. Este deslocamento é feito numa sucessão de passos cautelosos. Quando o avanço cessar de produzir melhores respostas, novo delineamento é estabelecido no sentido de se determinar a nova direção ótima de progressão. No caso de se verificar a inadequação do modelo planar, o delineamento de 1ª ordem é expandido para outro de 2ª ordem pela adição de novos pontos experimentais apropriadamente posicionados. Os dados produzidos permitirão o ajuste de um modelo de 2ª ordem.

Neste caso - na eventualidade do aparecimento de uma curvatura no sistema um polinômio de 2ª ordem é então aproximado

$$\eta = \alpha_0 + \sum_{i=1}^k \alpha_i \xi_i + \sum_{i=1}^k \sum_{j=1}^k \alpha_{ij} \xi_i \xi_j$$

Em geral é pouco provável que um modelo polinomial seja uma aproximação razoável da verdadeira relação funcional sobre todo o espaço das variáveis preditoras, mas para uma sub-região estreita desse espaço ele será frequentemente uma boa aproximação. Com esses objetivos experimentos são organizados para a produção de medidas adequadas e confiáveis da resposta de interesse em diversos pontos experimentais. A seguir o modelo matemático que melhor se ajusta aos dados é determinado através das técnicas estatísticas. Estes modelos apontarão as direções ótimas de avanço no espaço das variáveis preditoras. Aplicando sucessivos modelos locais sempre que a direção escolhida cessar de produzir ganhos na resposta, o conjunto das condições de operações ótimas para o sistema é finalmente determinado, estabelecendo-se a sub-região do espaço das variáveis preditoras que produz o valor ótimo da resposta.

No caso $k=1$, a relação funcional da variável resposta η com a variável preditora pode ser vista como uma curva de resposta, representada na figura 1.1. Já para o caso $k=2$, duas variáveis preditoras, a superfície de resposta será um subconjunto do R^3 como mostra a figura 1.2. No caso de $k \geq 3$, a superfície de respostas será uma superfície em um espaço de dimensão $k+1$ e o termo superfície de resposta continuará a ser usado para denominação da representação gráfica.

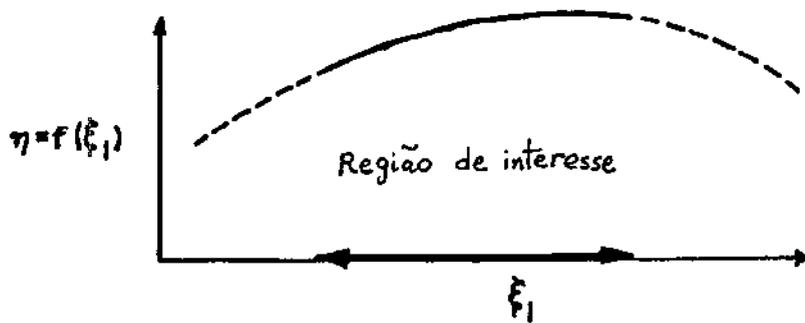


FIGURA 1.1 : Curva de Resposta

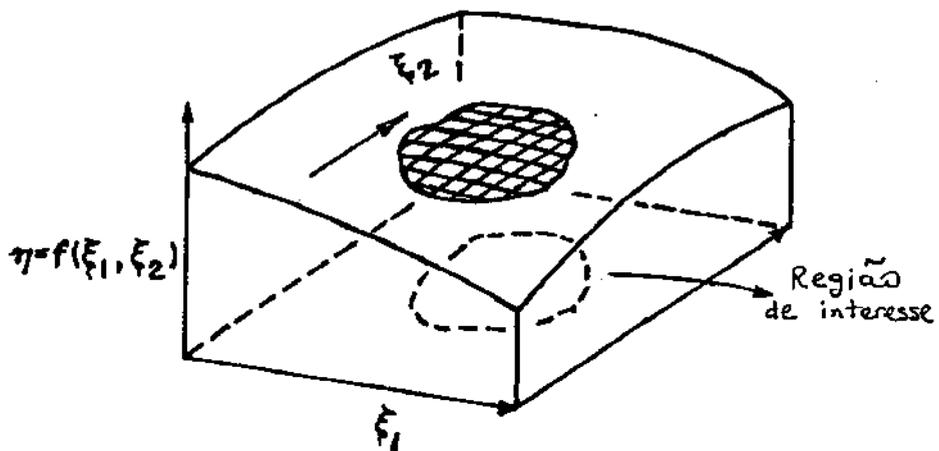


FIGURA 1.2 : Superfície de Respostas

1.2 TERMINOLOGIA EM MSR :

1.2.1 Fatores :

Os fatores são as variáveis preditoras que de alguma forma influenciam a resposta, ou seja, são as condições de operação que interagem no sistema em estudo. Assim, qualquer alteração feita no conjunto dos fatores será refletida na resposta.

O experimentador tem, por suposição, o controle sobre o conjunto dos fatores. Cada fator do sistema pode ou não ter vários níveis, que são por sua vez controlados (sem erros) pelo pesquisador. Por exemplo: temperatura de cozimento, concentração de reagente, velocidade de agitação, etc podem ser dispostos em uma escala numérica que é controlada pelo experimentador. Esses fatores e seus níveis serão denotados por $\xi_1, \xi_2, \dots, \xi_k$. O uso de variáveis preditoras na forma codificada é uma maneira de evitar problemas de escala como também facilitar o trabalho de cálculo nas análises. Para isso a padronização é feita por meio das variáveis $X_i = (\xi_i - \bar{\xi})/\delta(\xi_i)$, $i=1, \dots, k$, onde $\bar{\xi}$ é um fator de locação e $\delta(\xi_i)$ é um fator de escala. As variáveis preditoras assim codificadas denominamos X_1, X_2, \dots, X_k .

1.2.2 Resposta :

A resposta , η , é uma variável quantitativa que está sob influência direta dos fatores. Variação nos níveis dos fatores provocam alterações determinísticas no nível médio da resposta, as quais se sobrepõe o desvio aleatório médio esperado. A resposta para um determinado conjunto dos fatores é denotada por $\eta = f(\xi_1, \xi_2, \dots, \xi_k) = g(X_1, X_2, \dots, X_k)$. E na presença do erro experimental, ε , devido a erros operacionais de equipamentos, de medida, ou até mesmo a um fator aleatório do processo (comum em qualquer experimento), esta resposta (ou seja, o valor observado) ficará escrita como $y = \eta + \varepsilon$, com $E[\varepsilon] = 0$. Assim $\hat{y} = E[y] = E[\eta + \varepsilon] = \eta$.

1.2.3 A região operacional e a região experimental :

A região operacional (R_o) é a região onde é tecnicamente possível variar os níveis dos fatores. A região experimental (R_e) é um sub-conjunto da região operacional que é escolhido para um determinado experimento. Essa sub-região está totalmente contida na região operacional, como mostra a figura 1.2.3.1.

Devido à inviabilidade do estudo sobre toda a região operacional por questões próprias dos experimentos industriais o pesquisador é obrigado a fazer experimentos em sub-regiões pequenas da R_o . A MSR se ajusta naturalmente a estas restrições.

1.2.4 A representação gráfica :

A representação gráfica de uma superfície de respostas k -dimensional pode ser feita através de curvas de nível. O gráfico mostra o contorno da superfície para cada valor fixado da resposta nas diferentes combinações dos níveis dos fatores, como ilustrado na figura 1.2.4.1, onde estão representadas as curvas de níveis da superfície de respostas $\eta = 80 + 4X_1 + 8X_2 - 4X_1^2 - 12X_2^2 - 12X_1X_2$.

Quando a dimensão dos fatores, isto é, o número de fatores em estudo for $k \leq 3$, figura 1.2.4.1 as curvas de nível ainda são visualizáveis, quando $k \geq 4$ o indicado seria a fixação em um determinado nível de um dos fatores e então fazer a construção de um gráfico com apenas dois fatores.

Mais recentemente, com a popularização do uso de recursos computacionais gráficos, curvas que simulam uma visão tri-dimensional da superfície de respostas se tornaram comuns, como mostra a figura 1.2.4.2, representando a superfície de respostas $\eta = 79.94 + 0.9949X_1 + 0.5151X_2 - 1.377X_1^2 - 1.0018X_2^2 + 0.25X_1X_2$.

1.2.5 O ponto estacionário :

O ponto estacionário de uma superfície de resposta, y_0 , é o ponto onde a $\delta\eta/\delta x_i = 0$ para todo i . Este ponto pode corresponder a um máximo, um mínimo ou a uma sela, conforme a matriz das derivadas segundas seja negativa definida, positiva definida ou

indefinida.

Algumas vezes o ponto estacionário pode estar fora da região operacional. Nestes casos o mais indicado seria um novo direcionamento do estudo, se possível, para a região onde existam indícios de acréscimo (máximo) ou decréscimo (mínimo) na resposta.

As coordenadas do ponto estacionário do modelo com k variáveis preditoras codificadas serão denotadas por $X_{1s}, X_{2s}, \dots, X_{ks}$ que na forma natural se convertem a $\xi_{1s}, \xi_{2s}, \dots, \xi_{ks}$.

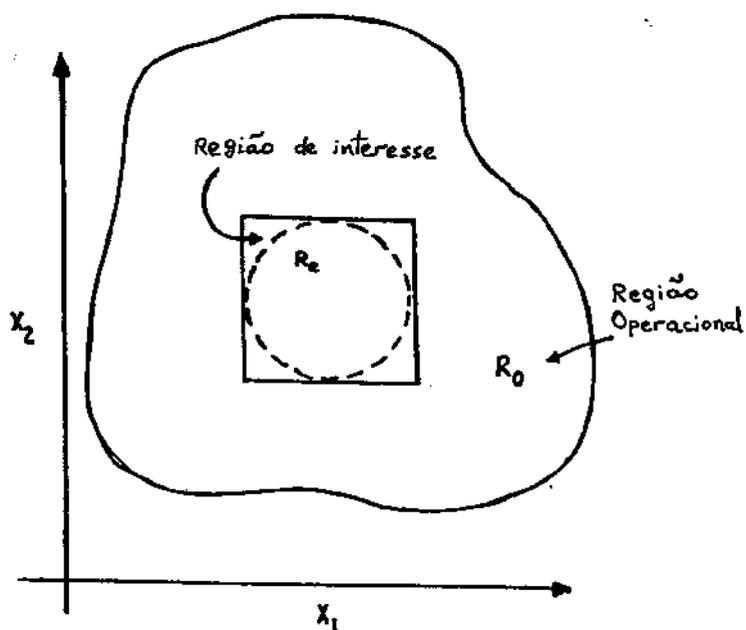


FIGURA 1.2.3.1 : Região operacional (R_0) e Região experimental (R_e) em uma região no R^2 .

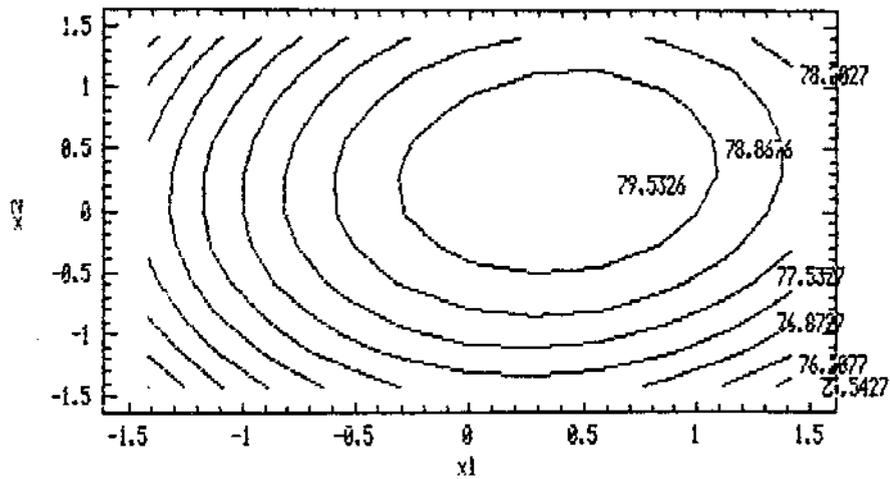


FIGURA 1.2.4.1 : Curvas de nível ajustada por um modelo polinomial de 2a. ordem com duas variáveis predictoras.

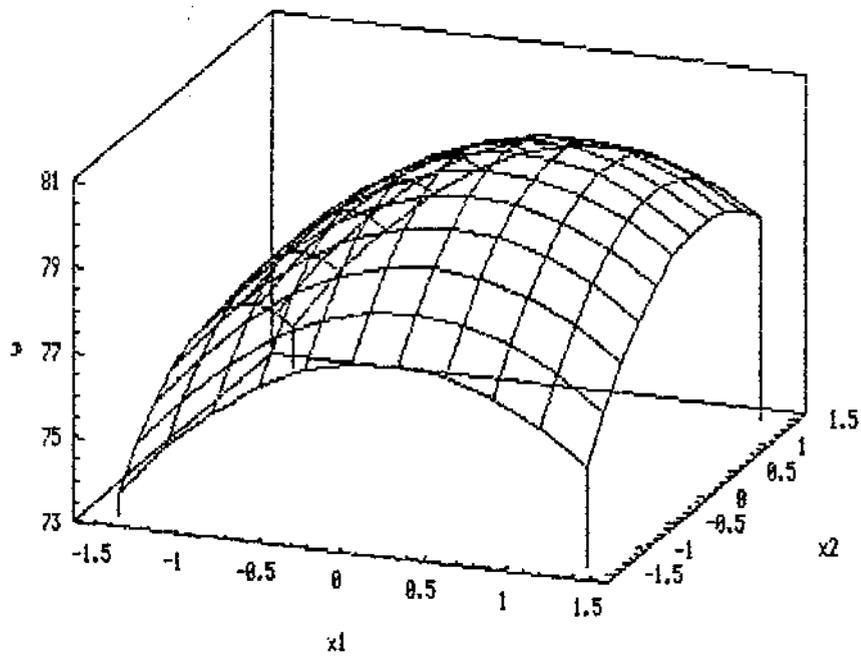


FIGURA 1.2.4.2 : Superfície de Respostas ajustada por um modelo polinomial de 2a. ordem com $k=2$.

1.3 APROXIMAÇÃO DA FUNÇÃO RESPOSTA :

1.3.1 O modelo de 1a. ordem :

A relação funcional (f) entre a resposta verdadeira (η) e o conjunto das variáveis preditoras $(\xi_1, \xi_2, \dots, \xi_k)$ é desconhecida quando se faz uma investigação para o ajuste de uma superfície de respostas. Através da utilização de modelos polinômiais, aproximações são feitas para a caracterização dessas funções em vizinhanças adequadamente pequenas.

Modelos polinômiais de ordens menores são mais utilizados na prática por causa do menor número de termos nos modelos. Eles são em geral suficientes porque os experimentos são realizados sobre sub-regiões estreitas na região operacional, denominada de região experimental (Re). Tipicamente a exploração de uma superfície de respostas se dá com a aplicação de modelos polinômiais de primeira e segunda ordem.

O modelo inicialmente utilizado em uma região experimental (Re) definida para um conjunto de k variáveis preditoras codificadas (X_1, X_2, \dots, X_k) é definido da seguinte forma:

$$y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \varepsilon \quad (1.3.1)$$

onde o erro experimental $\varepsilon \sim N(0, \sigma^2)$, independentes e identicamente

distribuídos (i.i.d.), $X_i = (\xi_i - \bar{\xi}) / \delta(\xi_i)$ são as variáveis preditoras na forma codificada, y é a variável resposta observada e os β_i 's, $i=1, \dots, k$, são os parâmetros desconhecidos do modelo e n o número de observações do experimento. Na forma matricial o modelo (1.3.1) fica escrito como:

$$\underset{\sim}{Y} = D \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad (1.3.2)$$

onde: $\underset{\sim}{Y}$: é o vetor $(n \times 1)$ de respostas observadas =
$$\begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$$

D : é a matriz $(n \times (k+1))$ dos níveis das variáveis preditoras codificadas denominada por matriz do delineamento,

$$D = \begin{bmatrix} 1 & x_{11} & \dots & \dots & x_{1k} \\ 1 & x_{21} & \dots & \dots & x_{2k} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & \dots & \dots & x_{nk} \end{bmatrix}$$

$\underset{\sim}{\beta}$: é o vetor $[(k+1) \times 1]$ dos parâmetros do modelo =
$$\begin{bmatrix} \beta_0 \\ \vdots \\ \vdots \\ \beta_k \end{bmatrix}$$

$\underset{\sim}{\varepsilon}$: é o vetor $(n \times 1)$ dos erros experimentais =
$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Os estimadores dos elementos de $\hat{\beta}$ são calculados pelo método de mínimos quadrados [16], pag. 70, que na forma matricial são dados por:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (D'D)^{-1} D' Y \quad (1.3.3)$$

onde $(D'D)^{-1}$ é a inversa da matriz simétrica $(D'D)$. A matriz de variância e covariância de $\hat{\beta}$ é dada por:

$$\text{Var } \langle \hat{\beta} \rangle = \text{Var } [(D'D)^{-1} D' Y] = (D'D)^{-1} \sigma^2$$

e o vetor de esperanças dado por:

$$E \langle \hat{\beta} \rangle = E [(D'D)^{-1} D' Y] = \beta \quad (1.3.5)$$

Assim, assumindo a normalidade dos erros, $\hat{\beta} \sim N(\beta ; (D'D)^{-1} \sigma^2)$.

1.3.1.1 Delineamentos para ajuste de modelos de 1a. ordem :

O ajuste e a análise de uma superfície de resposta fica muito mais facilitado com a escolha adequada de um delineamento experimental. Com a definição da região experimental para o ajuste de um modelo de primeira ordem em k variáveis preditoras do tipo (1.3.1), o passo seguinte é a escolha de um delineamento que facilite ao máximo a análise do modelo ajustado na região.

Entre os tipos de delineamentos para ajuste de modelos de primeira ordem, existe uma única classe de delineamentos que minimizam a variância dos coeficientes de regressão $\hat{\beta}_i$'s, $i \geq 1$. São os chamados delineamentos ortogonais de primeira ordem (ver [14], pag.478 e [4]). Um delineamento de primeira ordem é dito ortogonal quando os elementos fora da diagonal principal da matriz $(D'D)$ forem todos iguais a zero.

Na classe desses delineamentos estão incluídos os delineamentos fatoriais 2^k e os fatoriais 2^k fracionados. Para o uso destes delineamentos é feita uma codificação dos k fatores do fatorial nos níveis ± 1 , apesar de que desta maneira este delineamento não fornece uma estimativa do erro experimental e muito menos graus de liberdade para o teste da falta de ajuste do modelo, a não ser que hajam replicações no experimento. A introdução de replicações em alguns pontos do delineamento de primeira ordem pode resultar em um delineamento que tenha propriedades diferentes. Assim,

a propriedade de ortogonalidade do delineamento inicial poderia não permanecer invariante após as replicações. Se replicações forem feitas em todos os pontos do fatorial a ortogonalidade é mantida, contudo um procedimento destes pode acarretar um custo excessivo ao experimento. Assim, algumas replicações são incluídas nos pontos centrais ($X_i = 0, i=1, \dots, k$) para se obter uma estimativa do erro experimental. Esse aumento de observações nos pontos centrais não produz nenhum aumento nas estimativas dos coeficientes da regressão, $\hat{\beta}_i$'s, $i=1, \dots, k$, influenciando apenas a estimativa $\hat{\beta}_0$, que se torna uma média do número total de pontos do delineamento. Logo, não se altera a propriedade de ortogonalidade do delineamento. O teste da falta de ajuste do modelo pode ser realizado com base nos graus de liberdade provenientes das replicações no ponto central do delineamento, que nos permite uma estimativa pura do erro.

A matriz D de um modelo de primeira ordem em k variáveis fica escrita para os $n_f = 2^k$ pontos do fatorial e n_c pontos centrais como

$$D = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ 1 & -1 & -1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n_f 1} & \dots & x_{n_f k} \\ 1 & x_{n_f+1, 1} & \dots & x_{n_f+1, k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}$$

onde: $n = n_f + n_c$: é o número de observações no delineamento

$p = k+1$: é o número de parâmetros no modelo (1.3.1).

1.3.2 O modelo de 2ª ordem :

A falta de ajuste no modelo de primeira ordem (1.3.1), detectada pelo teste da falta de ajuste, é um indício de mudança na topografia da superfície. Assim, com o acréscimo de alguns pontos no delineamento onde foi detectada a falta de ajuste, este pode ser ampliado para um delineamento ortogonal de 2ª ordem, para o ajuste de um modelo de 2ª ordem

$$y = \beta_0 + \sum_{i=1}^k \beta_i X_i + \sum_{i=1}^k \sum_{j=1}^k \beta_{ij} X_i X_j + \varepsilon \quad (1.3.2.1)$$

onde: X_i : são as variáveis preditoras codificadas.

y : é a resposta que é influenciada pelo conjunto das variáveis preditoras.

$\beta_0, \beta_i, \beta_{ij}, (i,j=1, \dots, k)$: são os parâmetros desconhecidos do modelo.

ε : é o erro experimental associado à observação y .

Na forma matricial o modelo de segunda ordem (1.3.2.1)

fica escrito como

$$\underset{\sim}{Y} = \underset{\sim}{X} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \quad (1.3.2.2)$$

onde: $\underset{\sim}{Y}$: é o vetor (nx1) de respostas observadas = $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

D : é a matriz (nxp) dos valores das variáveis preditoras codificadas denominada por matriz do delineamento,

$$D = \begin{bmatrix} 1 & x_{11} & \dots & \dots & x_{1p} \\ 1 & x_{21} & \dots & \dots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & \dots & \dots & x_{np} \end{bmatrix}$$

$\underset{\sim}{\beta}_{(k)}$: é o vetor (px1) dos parâmetros desconhecidos do modelo = $\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{1k} \\ \vdots \\ \beta_{k-1,k} \end{bmatrix}$

$\underset{\sim}{\varepsilon}$: é o vetor (nx1) dos erros experimentais = $\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

p : é o número de parâmetros do modelo = (k + 1)(k + 2)/2.

Os estimadores de mínimos quadrados de $\underset{\sim}{\beta}_{(k)}$ no modelo (1.3.2.2) (ver [16], pag. 70) são dados por:

$$\underset{\sim}{\hat{\beta}}_{(k)} = (D'D)^{-1}D'Y \quad (1.3.2.3)$$

onde $(D'D)^{-1}$ é a inversa da matriz simétrica $(D'D)$. Assumindo a normalidade dos erros, com média 0 e desvio padrão σ , tem-se que:

$$\hat{\beta}_{(1)} \sim N(\beta_{(1)}; (D'D)^{-1} \sigma^2)$$

$$\hat{\beta}_i \sim N(\beta_i; (D'D)^{-1}_{ii} \sigma^2)$$

Assim, o modelo de segunda ordem ajustado na região do delineamento é dado por:

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i X_i + \sum_{i=1}^k \sum_{j=1}^k \hat{\beta}_{ij} X_i X_j \quad (1.3.2.4)$$

Com isso, o teste da falta de ajuste é realizado e o modelo (1.3.2.4) é então analisado. Não havendo falta de ajuste no modelo as coordenadas do ponto estacionário são então calculadas, $X_s = (X_{1s}, X_{2s}, \dots, X_{ks})$ como também o valor estimado da resposta no ponto estacionário \hat{y}_s . As coordenadas X_s são calculadas pela diferenciação da equação do modelo de segunda ordem (1.3.2.4) com relação a cada X_i . O modelo (1.3.2.4) pode ser escrito na seguinte forma matricial

$$\hat{Y} = \hat{\beta}_0 + X' \hat{\beta} + X' B X \quad (1.3.2.5)$$

onde:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \quad e$$

$$B = \begin{pmatrix} \hat{\beta}_{11} & & & \\ \hat{\beta}_{12}/2 & \hat{\beta}_{22} & & \text{sim.} \\ \vdots & \vdots & \ddots & \\ \hat{\beta}_{1k}/2 & \hat{\beta}_{2k}/2 & \dots & \hat{\beta}_{kk} \end{pmatrix}$$

As derivadas parciais de (1.3.2.5) com relação a cada X_i são dadas por:

$$\left. \begin{aligned} \frac{\partial \hat{y}}{\partial X_1} &= \hat{\beta}_1 + 2\hat{\beta}_{11} X_1 + \sum_{i:1}^k \hat{\beta}_{1i} X_i \\ \frac{\partial \hat{y}}{\partial X_2} &= \hat{\beta}_2 + 2\hat{\beta}_{22} X_2 + \sum_{\substack{i:1 \\ i \neq 2}}^k \hat{\beta}_{2i} X_i \\ &\vdots \\ \frac{\partial \hat{y}}{\partial X_k} &= \hat{\beta}_k + 2\hat{\beta}_{kk} X_k + \sum_{i:1}^{k-1} \hat{\beta}_{ki} X_i \end{aligned} \right\} = \hat{\beta} + 2BX$$

Igualando-se cada uma das derivadas a zero e assumindo que B é não singular, tem-se que:

$$X_s = -B^{-1} \hat{\beta}/2 \quad (1.3.2.6)$$

onde B^{-1} é a inversa da matriz simétrica B. Substituindo-se X_s na equação (1.3.2.5), o valor da resposta predita no ponto estacionário é dado por

$$\hat{y}_s = \hat{\beta}_0 + X_s' \hat{\beta}/2$$

1.3.2.1 Delineamentos rotacionáveis para modelos de 2ª ordem :

Os delineamentos construídos para o ajuste de modelos do tipo (1.3.2.1) são chamados de delineamentos de 2ª ordem. Para tais ajustes o número de pontos no delineamento tem que ser pelo menos igual ao número de parâmetros no modelo, ou seja, $n \geq p$, onde $p = (k+1)(k+2)/2$. Por exemplo, para $k=2, 3$, e 4 , p seria igual a $6, 10$ e 15 , respectivamente. O ideal é tentar conciliar um número mínimo de pontos experimentais que sejam suficientes para a estimação dos parâmetros do modelo (1.3.2.1) com uma confiança aceitável.

Na literatura referente ao assunto, existem diversos delineamentos que podem ser utilizados para o ajuste de modelos de 2ª ordem. O objetivo então é o uso de algum critério para a escolha do tipo de delineamento a ser usado. No caso dos delineamentos de 1ª ordem, a ortogonalidade é um bom critério, pois otimiza a variância dos coeficientes dos estimadores da regressão.

Box e Hunter [3] em 1957 publicaram um artigo ressaltando propriedades desejáveis em delineamentos de 2ª ordem, entre elas a ortogonalidade e rotacionalidade. Um delineamento de 2ª ordem é dito ortogonal quando as colunas da matriz de delineamento X , que se referem aos níveis das variáveis $X_1, X_2, \dots, X_k; X_1^2, \dots, X_k^2; X_1X_2, \dots, X_{k-1}X_k$, puderem ser escritas como polinômios ortogonais, isto é, $\sum_{i=1}^n X_{li}X_{lj} = 0, \quad i, j = 1, 2, \dots, (2k+1), \quad i \neq j$, como mostra a matriz abaixo para um caso onde $k=2$.

$$\begin{array}{cccccc}
 & 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1 X_2 \\
 & \hline
 & 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1 X_2 \\
 & \hline
 D = & \begin{pmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -\sqrt{2} & 0 & 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 & 2 & 0 & 0 \\ 1 & 0 & -\sqrt{2} & 0 & 2 & 0 \\ 1 & 0 & \sqrt{2} & 0 & 2 & 0 \end{pmatrix} & \rightarrow & (D'D) = & \begin{pmatrix} 10 & 0 & 0 & 8 & 8 & 0 \\ 0 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 \\ 8 & 0 & 0 & 12 & 4 & 0 \\ 8 & 0 & 0 & 4 & 12 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}
 \end{array}$$

O delineamento de 2ª ordem é dito rotacionável se o valor da variância da resposta predita, \hat{y} , permanecer constante para os pontos equidistantes do centro do delineamento (ver [3], pag. 213 a 215). Isto faz com que a $\text{Var}(\hat{y})$ seja função apenas da distância do centro do delineamento e não da direção. Esta é uma importante propriedade, pois, sendo o objetivo na MSR a otimização das condições de operação e a localização do ótimo é desconhecida antes da realização do experimento, é razoável a utilização de delineamentos que produzam igual precisão da estimativa da resposta em todas as direções.

Uma classe de delineamentos bastante utilizada para o ajuste de modelos de 2ª ordem são os delineamentos compostos centrais, introduzidos no meio científico por Box e Wilson (ver [1]) em 1951. Este delineamento utiliza 2^k pontos do fatorial e n_c pontos centrais com a adição de $2k$ pontos axiais, chamados de pontos da estrela, como mostra a figura 1.3.2.2 para delineamentos com 2 e 3 variáveis. Tal

delineamento mostrou ter a vantagem de poder ser executado em estágios, visto que o delineamento de 1ª ordem poderia ser aproveitado com a introdução dos pontos axiais para o ajuste do modelo de 2ª ordem, além da utilização de um número menor de pontos experimentais se comparado com um fatorial completo para o ajuste de 2ª ordem. O delineamento de Box também mostrou ter as propriedades de ortogonalidade e rotacionalidade (ver [3]).

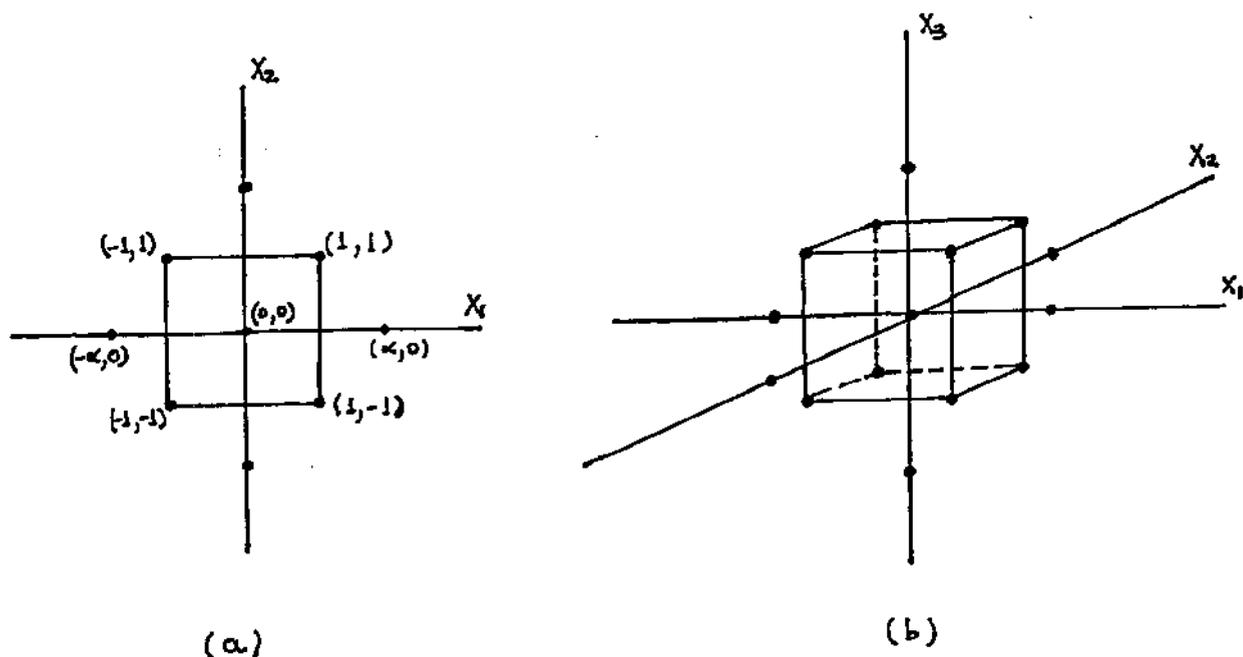


FIGURA 1. 3. 2. 2 : Delineamento Composto Central (a) com $k=2$ e (b) $k=3$ com 2^k pontos no fatorial, n_c pontos centrais e $2k$ pontos na estrela.

1.4 O METODO DO " STEEPEST ASCENT " :

O Método de Exploração de Superfície de Resposta é um procedimento baseado em experimentos sequenciados em pequenas regiões. De acordo com os resultados nestes experimentos iniciais outros delineamentos são encaminhados sucessivamente até que as condições ótimas de operação sejam encontradas, o que equivale à determinação do ponto estacionário da superfície.

O ajuste de uma função sobre toda a região operacional seria impossibilitada pelo número excessivo de experimentos que teriam que ser conduzidos. Para tal ajuste, um método de aproximação é utilizado em uma pequena região experimental ($R_e \subset R_o$) de interesse do pesquisador. Tais aproximações são feitas por modelos polinômiais. Nestas sub-regiões a topografia da superfície ajustada não é necessariamente a mesma, ou talvez em nada se assemelhe à topologia da superfície verdadeira. Assim a aproximação feita através do polinômio somente é útil na vizinhança da região experimental onde ela foi ajustada. Para isso, inicialmente são feitos ajustes de polinômios de primeira ordem do tipo

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i X_i$$

Com o eventual aparecimento de uma curvatura no sistema, um polinômio de grau maior tem que ser usado, tal como um polinômio de

segundo grau

$$\hat{y} = \hat{\beta}_0 + \sum_{l=1}^k \hat{\beta}_l X_l + \sum_{l=1}^k \sum_{j=1}^k \hat{\beta}_{lj} X_l X_j .$$

Na opinião de muitos autores, o método da direção mais íngreme de percurso (*steepest ascent*), introduzido na literatura por Box e Wilson [1] em 1951 tem se mostrado um método preliminar bastante eficaz e económico. Este procedimento permite um ajuste de segunda ordem após alguns ajustes iniciais de primeira ordem. O método permite a escalada da superfície através do gradiente da função do ajuste de primeira ordem (ver [1]). A direção do *steepest ascent* é assim determinada e um novo delineamento é então direccionado até que a falta de ajuste do modelo de primeira ordem seja detectada, o que indica uma mudança na curvatura da superfície e aproximação da vizinhança do ponto estacionário. Quando ocorre esta falta de ajuste no modelo polinomial de primeira ordem, alguns pontos são adicionados ao delineamento e então um novo modelo polinomial de segunda ordem é ajustado e o teste da falta de ajuste é novamente realizado. Este processo se repete até a chegada na região de vizinhança do ponto estacionário da superfície.

1.5 CARACTERIZAÇÃO DA SUPERFÍCIE DE RESPOSTAS :

Após a determinação do ponto estacionário da superfície, o interesse passa a ser a caracterização da natureza da superfície na vizinhança da estimativa do ponto estacionário

$$\hat{y}_e = \hat{\beta}_0 + X_e' \hat{\beta} / 2 \quad (1.5.1)$$

com $X_e = - B^{-1} \hat{\beta} / 2$, o que equivale à determinação do ponto ótimo de operação. A caracterização da superfície se dá pela identificação do ponto estacionário ser um ponto de máxima ou mínima resposta ou um ponto de sela, e também pela sensibilidade da resposta com relação às variáveis preditoras na vizinhança do ponto estacionário.

Quando se trabalha em um sistema com no máximo três variáveis preditoras ($k \leq 3$), o indicado é um estudo gráfico das curvas de nível do modelo ajustado. Contudo, se $k > 3$ um estudo deste tipo se torna inviável pelas dimensões do espaço das variáveis preditoras. Para isso (e até mesmo quando se tem poucas variáveis preditoras) e para uma análise mais detalhada do modelo de segunda ordem ajustado, o uso de uma Análise Canônica é feito para uma caracterização mais formal e uma melhor interpretação da superfície.

Análise Canônica é um método que consiste em uma mudança de eixos do sistema com o novo centro do sistema sendo o ponto estacionário, o que causa a remoção dos termos de primeira ordem

$(\hat{\beta}_i, i=1, \dots, k)$, seguida de uma rotação de eixos que remove os termos dos produtos cruzados do modelo de segunda ordem $(\hat{\beta}_{ij}; i, j=1, \dots, k; i \geq j)$, assim o sistema se reduz a um conjunto de variáveis canônicas denotadas por W_1, W_2, \dots, W_k , como mostra a figura 1.5.1 com $k=2$. A seguir são mostrados os passos para esta transformação.

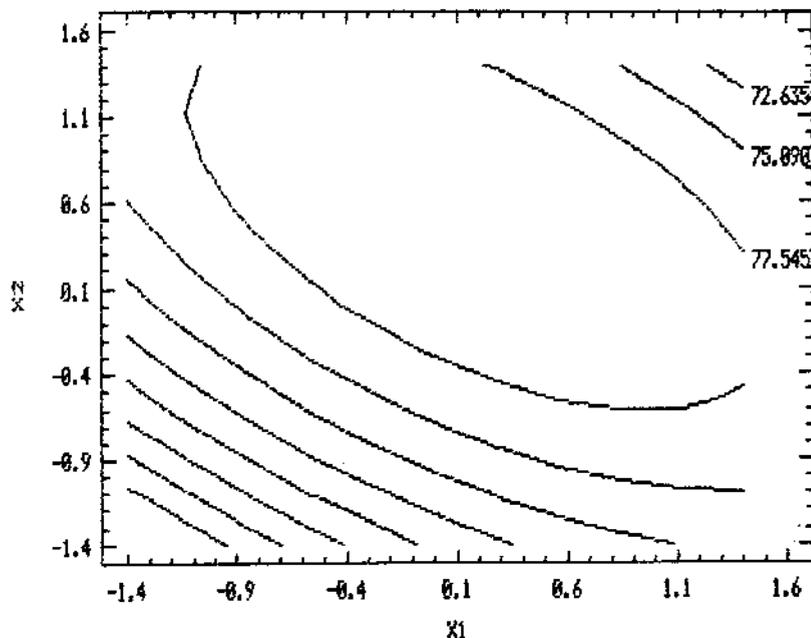


FIGURA 1.5.1 : Equação Canônica de um modelo com $k=2$,
 $\hat{y} = 80 - 4W_1^2 - W_2^2$, onde S é o centro
do sistema (W_1, W_2) .

(1) O primeiro passo é feito com a translação da origem do sistema para o ponto estacionário através da transformação

$$\tilde{X} = (X - X_s)$$

onde: $X = (X_1, X_2, \dots, X_k) = (0, 0, \dots, 0)$ são as coordenadas do centro do sistema antigo.

$X_s = (X_{1s}, X_{2s}, \dots, X_{ks})$ são as coordenadas do ponto estacionário do delineamento onde foi ajustado o modelo de segunda ordem com a solução $X_s = -B^{-1} \hat{\beta}/2$.

$\tilde{X} = (X_1 - X_{1s}, X_2 - X_{2s}, \dots, X_k - X_{ks})$ são as coordenadas do centro do novo sistema.

Fazendo esta transformação no modelo de segunda ordem ajustado têm-se que:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + X' \hat{\beta} + X' B X = \\ &= \hat{\beta}_0 + (\tilde{X} + X_s)' \hat{\beta} + (\tilde{X} + X_s)' B (\tilde{X} + X_s) = \\ &= \hat{\beta}_0 + \tilde{X}' \hat{\beta} + X_s' \hat{\beta} + \tilde{X}' B \tilde{X} + X_s' B \tilde{X} + \tilde{X}' B X_s + X_s' B X_s = \\ &= \hat{\beta}_0 + \tilde{X}' \hat{\beta} + X_s' \hat{\beta} + \tilde{X}' B \tilde{X} - \hat{\beta}' B^{-1} B \tilde{X}/2 - \tilde{X}' B B^{-1} \hat{\beta}/2 - X_s' B B^{-1} \hat{\beta} \end{aligned}$$

como $B B^{-1} = I$, então:

$$\begin{aligned} \hat{Y} &= \hat{\beta}_0 + \tilde{X}' \hat{\beta} + X_s' \hat{\beta} + \tilde{X}' B \tilde{X} - \hat{\beta}' \tilde{X}/2 - \tilde{X}' \hat{\beta}'/2 - X_s' \hat{\beta} \\ \hat{Y} &= \hat{\beta}_0 + X_s' \hat{\beta}/2 + \tilde{X}' B \tilde{X} \end{aligned}$$

sendo $\hat{y}_s = \hat{\beta}_0 + X_s' \hat{\beta}/2$, tem-se que:

$$\hat{y} = \hat{y}_s + \tilde{X}' B \tilde{X} \tag{1.5.2}$$

onde: \hat{Y} : é o vetor (n x 1) dos valores preditos.

\hat{y}_s : é o valor da resposta estimada no ponto estacionário do modelo.

B : é a matriz simétrica (k x k) das estimativas dos termos de segunda ordem do modelo.

(ii) O segundo e último passo se dá com a rotação dos eixos ($\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_k$), removendo-se assim os termos do produto cruzado.

Essa rotação é feita em função de uma Decomposição Spectral realizada em B (ver [7], pag. 469). Pela transformação

$$B = M \Lambda M' = \sum_{i=1}^k \lambda_i m_{(i)} m_{(i)}', \text{ onde } \lambda_i \text{ e } m_{(i)} \text{ são os correspondentes}$$

autovalores e autovetores normalizados associados à matriz B, o que resulta em

$$\begin{aligned} \hat{y} &= \hat{y}_s + \tilde{X}' B \tilde{X} = \hat{y}_s + \tilde{X}' M \Delta M' \tilde{X} = \\ &= \hat{y}_s + W' \Delta W \end{aligned} \quad (1.5.3)$$

onde: $\Delta = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$

$$M = [m_{(1)} \quad m_{(2)} \quad \dots \quad m_{(k)}]$$

$$W = M' \tilde{X} = M' (X - X_s)$$

$$W_i = m_{(i)}' (X_i - X_{is}), \quad i=1, \dots, k.$$

Assim, o modelo ajustado depois da transformação fica escrito em função das variáveis canônicas como

$$\hat{y} = \hat{y}_s + \lambda_1 W_1^2 + \lambda_2 W_2^2 + \dots + \lambda_k W_k^2 \quad (1.5.4)$$

onde os λ_i 's são os autovalores da matriz B e representam os coeficientes dos termos W_i^2 's na equação canônica.

A equação canônica (1.5.4) do modelo de segunda ordem ajustado para uma superfície de resposta tem a mesma função do polinômio de segunda ordem, isto é, estimar a função da resposta verdadeira η em uma região restrita, com a vantagem da facilidade na interpretação dos coeficientes da equação canônica, ou seja, os λ_i 's. O sinal e a magnitude destes valores estão diretamente relacionadas com a topologia da superfície. Se $\lambda_1, \lambda_2, \dots, \lambda_k$ são:

- (i) Todos negativos, então no ponto estacionário X_s , a superfície tem um ponto de máximo (figura 1.5.2 a).
- (ii) Todos positivos, então no ponto estacionário X_s , a superfície possui um ponto de mínimo (figura 1.5.2 b).
- (iii) Alguns positivos e alguns negativos, então X_s é um ponto de sela (figura 1.5.2 c).

A magnitude dos valores absolutos dos λ_i 's, ou seja, $|\lambda_i|$, fornecem uma ideia de quanto a resposta é sensível em relação a direção de W_i a partir de X_s . Por exemplo, se $k=2$ e $|\lambda_1| > |\lambda_2|$, o valor da resposta a partir de X_s muda mais rapidamente quando a movimentação for feita na direção do eixo W_1 , figura 1.5.2 a. Quando um ou mais dos λ_i 's estão próximos de zero ou são iguais a zero, o sistema é insensível à variável W_i , isto é, não ocorrem mudanças relevantes na resposta se houver movimentação ao longo de W_i a partir de X_s (ponto de origem do sistema W_1, W_2, \dots, W_k e coordenadas da

estimativa do ponto estacionário da superfície). Para esses casos a superfície ajustada é denominada "stationary ridge" que fornece uma idéia de uma cordilheira no caso de $k=2$, como é mostrado na figura 1.5.3.

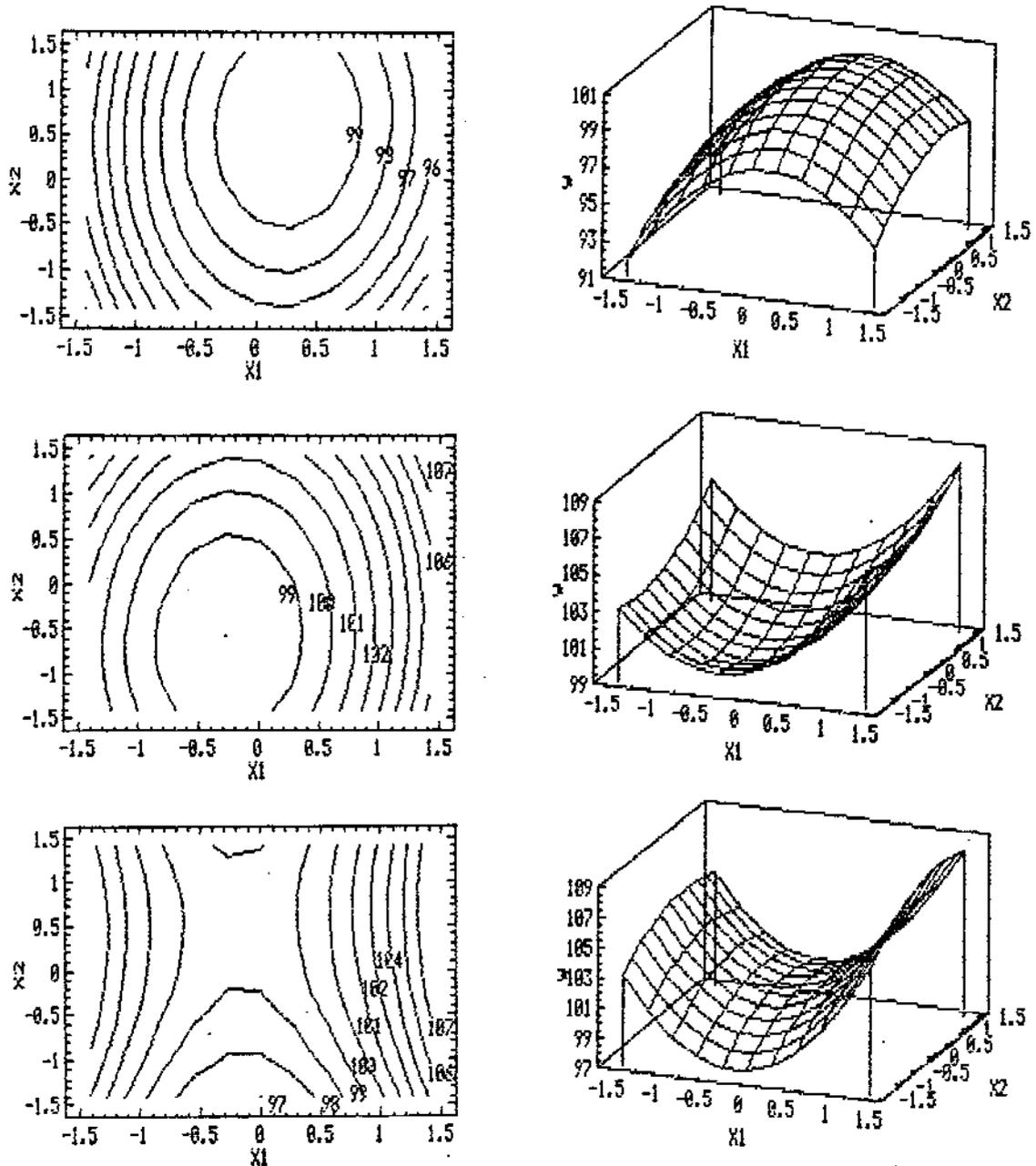


FIGURA 1.5.2 : Exemplos de Superfícies de respostas e curvas de níveis ajustadas por modelos de 2a. ordem com duas variáveis preditoras ($k=2$).

A utilização da relação entre as variáveis canônicas W_i e as variáveis do delineamento X_i pode ser necessária em algumas situações. Principalmente nos casos onde existe a impossibilidade de atuação do processo no ponto estacionário através das variáveis do delineamento, que pode ser causada por fatores como um custo excessivo, ou mesmo por problemas de mecanismo do sistema. Assim, através das variáveis canônicas W_i , sabe-se em que direção existe uma maior (ou menor) variação na resposta a partir do ponto estacionário em X_0 , podendo desta maneira haver uma conversão dos pontos no espaço de (W_1, W_2, \dots, W_k) para os pontos no espaço de (X_1, X_2, \dots, X_k) por meio da transformação $W = M'(X - X_0)$.

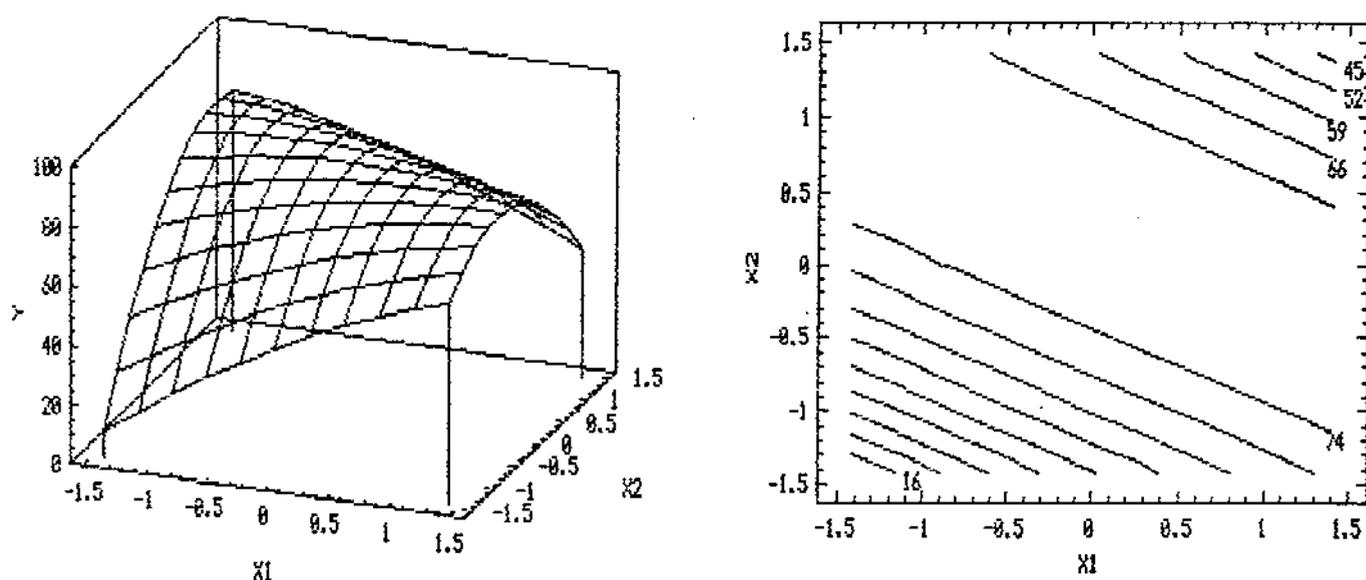


FIGURA 1.5.3 : Superfície de respostas e as curvas de nível ajustadas por um modelo polinomial de 2a. ordem, com duas variáveis preditoras, X_1 e X_2 , dado por $\eta = 80 + 4X_1 + 8X_2 - 3X_1^2 - 12X_2^2 - 12X_1X_2$.

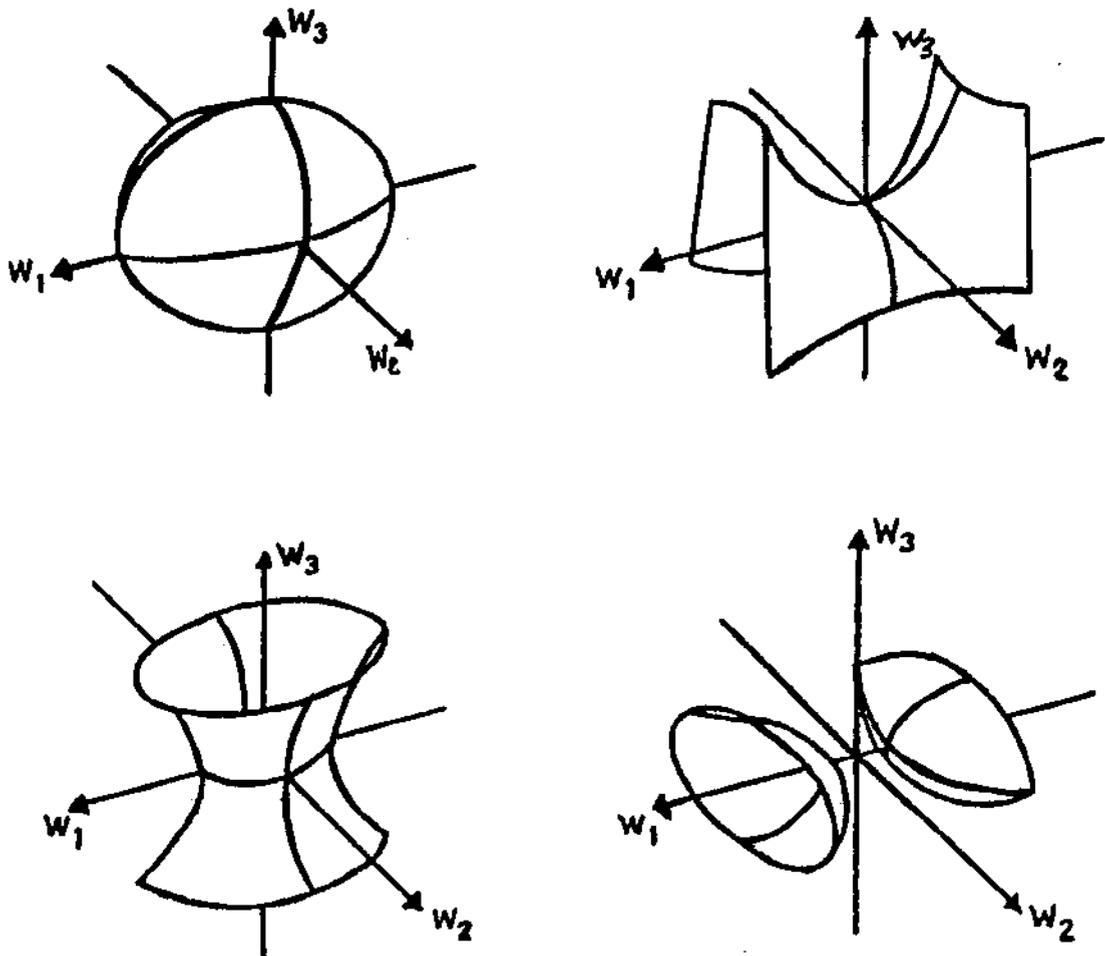


FIGURA 1.5.4 : Exemplos de curvas de níveis de superfície de respostas, ajustadas por modelos polinomiais de 2a. ordem, em função das variáveis canônicas W_1 , W_2 e W_3 .

CAPITULO 2 - O "BOOTSTRAP", UMA VISÃO GERAL E APLICAÇÃO NO CONTEXTO DE SUPERFÍCIE DE RESPOSTAS

2.1 O MÉTODO "BOOTSTRAP" :

O *bootstrap* é uma técnica de reamostragem relacionada com o *jackknife*, introduzida por Efron [6] em 1979, despertando desde então enorme interesse, tanto do ponto de vista da fundamentação e desenvolvimento teóricos, quanto das aplicações práticas. É um método computacionalmente intensivo e por isso o seu desenvolvimento e aplicação têm sido potenciados pelo avanço e popularização dos recursos computacionais.

O método pode ser considerado como uma ferramenta para soluções numéricas de diversos problemas estatísticos que muitas vezes são analiticamente inviáveis, tanto em contextos paramétricos como não-paramétricos. O método se baseia na reutilização da amostra X_1, X_2, \dots, X_n para a obtenção de estimativas da distribuição amostral de uma estatística de interesse $T(X_1, X_2, \dots, X_n)$.

Sejam X_1, X_2, \dots, X_n variáveis aleatórias independentes e identicamente distribuídas, com função distribuição F , desconhecida ou conhecida a menos de um ou mais parâmetros. Sejam ainda $(\hat{T}(X) = \hat{T}(X_1, X_2, \dots, X_n))$ estatísticas reais e $T(X)$ funcionais em F , tais que $(\hat{T}(X) - T(X))\sqrt{n} \xrightarrow{D} N(0, \text{Var}(T(X)))$, quando $n \rightarrow \infty$. Seja $H(X, F)$ a função de distribuição exata de $(\hat{T}(X) - T(X))\sqrt{n}$. Na maioria das vezes o grande problema na estatística é a estimação de $H(X, F)$ ou funcionais de $H(X, F)$ da amostra. O *bootstrap* estima $H(X, \hat{F})$ para $H(X, F)$, onde \hat{F} é a função de distribuição amostral. A idéia básica do *bootstrap* se dá na seguinte forma:

(1) Seja uma amostra aleatória de tamanho n de uma distribuição de probabilidade F , desconhecida,

$$\tilde{X} = (X_1, X_2, \dots, X_n) \sim F, \text{ i. i. d.}$$

tendo como valores observados $X_1 = x_1, \dots, X_n = x_n$ e $\tilde{x} = (x_1, \dots, x_n)$.

(ii) Dada uma determinada variável aleatória de interesse $T(\underline{X})$, o interesse é a estimação da distribuição amostral de T com base nos valores observados de $\underline{X} = \underline{x}$.

(iii) Constrói-se a distribuição de probabilidade amostral \hat{F} , colocando-se massa $1/n$ em cada ponto $x_i, i=1, \dots, n$.

(iv) Fixada \hat{F} , retira-se uma amostra aleatória com reposição de tamanho n de \hat{F} , denominada de

$$X_i^* = x_i^*, X_i^* \sim \hat{F} \text{ i.i.d.}, i=1, \dots, n.$$

Chamando-se assim $\underline{X}_{\sim}^* = (X_1^*, \dots, X_n^*)$ de uma amostra *bootstrap*.

(v) Aproximando-se assim a distribuição de $H(\underline{X}, \hat{F})$ pela distribuição *bootstrap* de $H(\underline{X}_{\sim}^*, \hat{F})$.

Uma estimativa da distribuição real *bootstrap* da estatística pode ser obtida via Monte Carlo. Repetidas realizações de \underline{X}_{\sim}^* são feitas, retirando-se amostras aleatórias simples com reposição, de tamanho n, de \hat{F} , ou seja, $\underline{X}_{\sim_1}^*, \underline{X}_{\sim_2}^*, \dots, \underline{X}_{\sim_{BS}}^*$, onde BS é o número de amostras *bootstrap*. O histograma para os valores correspondentes de $H(\underline{X}_{\sim_1}^*, \hat{F}), H(\underline{X}_{\sim_2}^*, \hat{F}), \dots, H(\underline{X}_{\sim_{BS}}^*, \hat{F})$ é construído como uma aproximação da distribuição real *bootstrap* da estatística e fornece uma visualização da forma da distribuição de $H(\underline{X}, \hat{F})$.

2.1.1 Exemplo :

O conjunto de dados bivariados utilizados neste exemplo, provém de DIACONIS e EFRON [11] que ilustra o uso do *bootstrap*. Os dados da tabela 2.1.1 se referem a duas medidas de potencial acadêmico (GPA e LSAT) de 15 Escolas Americanas de Direito que foram tomadas como amostra. O GPA é uma média classificatória universitária e LSAT é uma média obtida no Teste de Admissão nas Escolas de Direito. O coeficiente de correlação, ρ , foi estimado pelo coeficiente de correlação amostral, $\hat{\rho}$, dado por:

$$\hat{\rho} = \frac{\sum_{i=1}^{15} (X_i - \bar{X})(Y_i - \bar{Y})}{\left\{ \sum_{i=1}^{15} (X_i - \bar{X})^2 \sum_{i=1}^{15} (Y_i - \bar{Y})^2 \right\}^{1/2}} = 0.776$$

A estimativa *bootstrap* da distribuição de $\hat{\rho}$ com base em BS = 1000 repetições *bootstrap* apresentou média $\bar{\rho}^* = 0.787$ e um desvio padrão de 0.133 . O histograma das observações *bootstrap* $\hat{\rho}_1^*, \dots, \hat{\rho}_{nbs}^*$ de $\hat{\rho}$ está na figura 2.2.2. Este histograma se constitui na estimativa *bootstrap* da distribuição de $\hat{\rho}$.

tabela 2.1.1 : GPA e LSAT das 15 Escolas de Direito

GPA(X)	576	635	558	578	666	580	555	661	651	605	659	575
LSAT(Y)	3.39	3.30	2.81	3.09	3.44	3.07	3.00	3.43	3.36	3.19	3.12	2.74

cont. tabela 2.1.1

GPA(X)	544	572	594
LSAT(Y)	2.76	2.88	2.96

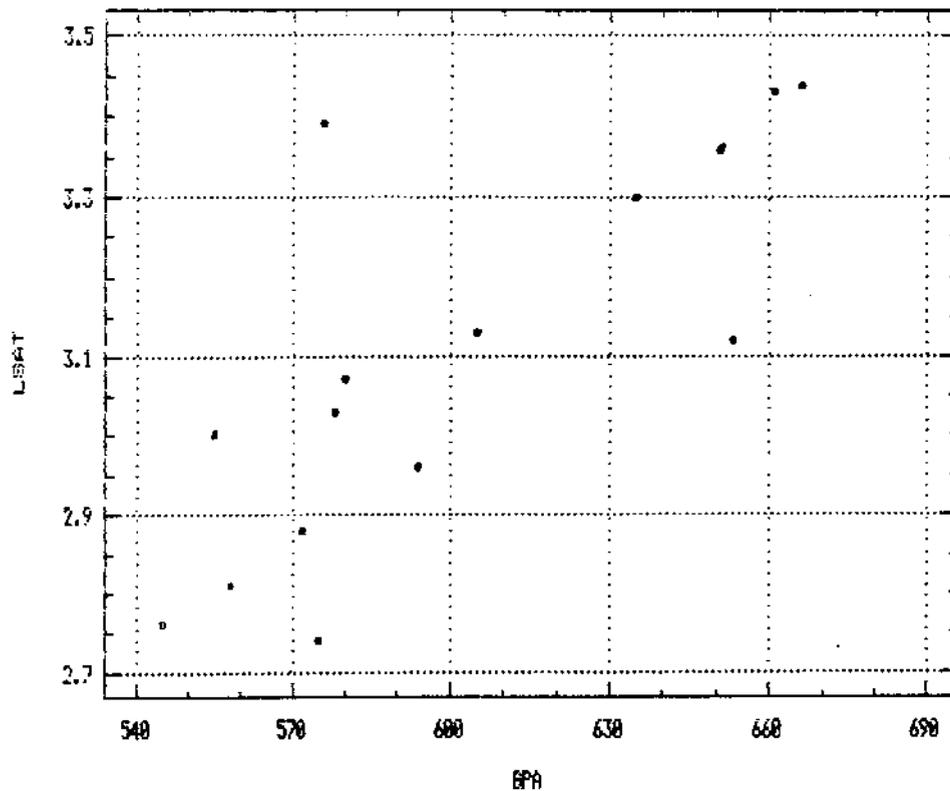


FIGURA 2.2.1 : Coeficiente de correlação amostral, $\hat{\rho}$, entre as medidas de GPA e LSAT das 15 escolas.

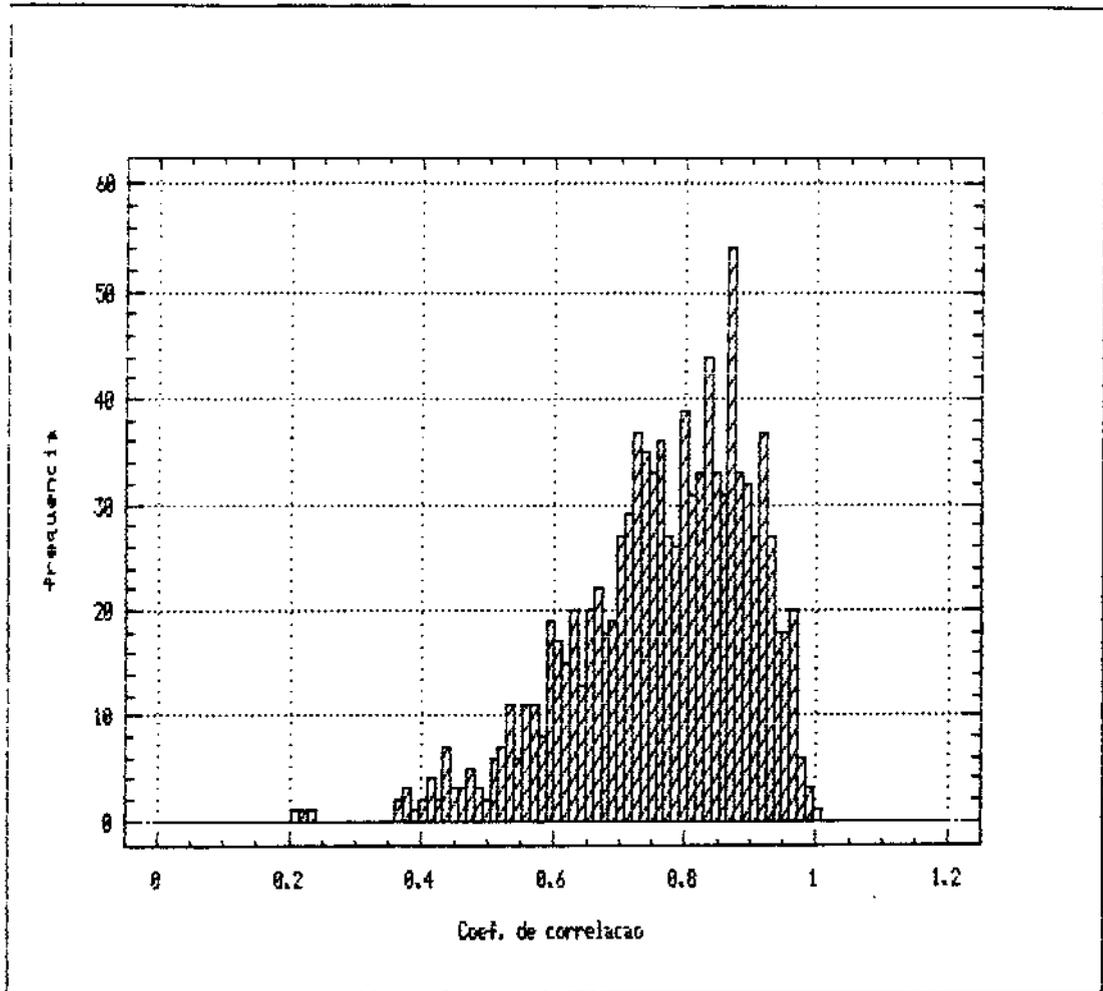


FIGURA 2.2.2 : Histograma com base em 1000 replicações bootstrap de $\hat{\rho}$ entre GPA e LSAT das 15 escolas de direito.

2.2 O "BOOTSTRAP" EM REGRESSÃO LINEAR:

A aplicação do método *bootstrap* em modelos de regressão teve início com Efron [6] em 1979. A partir de então vários outros autores como Freedman [8], Freedman e Peters [12], Chaves [17] e outros têm se empenhado no estudo da aplicação do método *bootstrap* em modelos de regressão.

A aplicação do método se dá, usualmente, para a obtenção da estimativa *bootstrap* da distribuição do vetor das estimativas dos parâmetros, $\hat{\beta}$, do modelo de regressão do tipo $\underline{Y} = D\beta + \underline{\varepsilon}$. Para essa aplicação os passos são os seguintes:

(i) Seja o modelo $\underline{Y} = D\beta + \underline{\varepsilon}$ onde \underline{Y} é o vetor $(n \times 1)$ de respostas observadas e $\underline{\varepsilon}$ o vetor $(n \times 1)$ de erros experimentais, β é o vetor $((p+1) \times 1)$ de parâmetros do modelo e D é a matriz fixa do delineamento de ordem $(n \times (p+1))$, sendo p o número de parâmetros do modelo;

(ii) Os erros são variáveis aleatórias independentes e identicamente distribuídas (i.i.d) com função distribuição desconhecida, F , de tal modo que $e_i \sim F$, i.i.d, $i=1, \dots, n$, onde assume-se que F é centrada de alguma forma em zero, com $E_F(e_i) = 0$ ou $\text{Mediana}_F(e_i) = 0$.

(iii) Assumindo-se que a matriz do delineamento, D , é fixa e sendo \underline{Y} o

vetor de respostas observadas, a estimação de $\hat{\beta}$ pode ser feita por meio do método de mínimos quadrados, que fornece o vetor de estimativas dos parâmetros $\hat{\beta} = (D'D)^{-1} D' Y$.

(iv) O vetor dos resíduos, \tilde{r} , é então determinado por:

$$\tilde{r} = Y - \hat{Y} = Y - D\hat{\beta}$$

(v) Os resíduos são multiplicados pelo fator $(n/(n-p))^{1/2}$ para compensar a redução na variabilidade associada à perda dos graus de liberdade, ficando assim $r_i = \tilde{r}_i (n/(n-p))^{1/2}$.

(vi) Uma distribuição empírica, \hat{F} , é assumida para os resíduos, \tilde{r} , associando a probabilidade $1/n$ para cada r_i , fazendo assim $r_i \sim \hat{F}$, i.i.d.

(vii) Retirando-se uma amostra aleatória simples com reposição, de tamanho n , de \hat{F} , denominada de amostra *bootstrap*

$$\tilde{r}^* = (r_1^*, r_2^*, \dots, r_n^*)$$

(viii) Calcula-se um vetor de pseudo-dados, \tilde{Y}^* , como sendo

$$\tilde{Y}^* = \hat{Y} + \varepsilon^* = D\hat{\beta} + \varepsilon^*$$

(ix) Ajustando-se novamente um modelo através de mínimos quadrados e obtendo-se a reestimativa *bootstrap* de $\hat{\beta}$ como sendo

$$\hat{\beta}^* = (D'D)^{-1} D' \tilde{Y}^*$$

Assim, via Monte Carlo, os passos (vi)-(viii) são repetidos para um número consideravelmente grande de vezes nbs , produzindo $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, ..., $\hat{\beta}_{nbs}^*$ como sendo uma aproximação da distribuição amostral de $\hat{\beta}$.

Os valores do *bootstrap*, r_i^* , reamostrados em (vi) são independentes com média zero e variância $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / n$. Significando assim que a estimativa *bootstrap*, $\hat{\beta}^*$, tem média e variância *bootstrap* dadas por:

$$\begin{aligned} E(\hat{\beta}^*) &= E((D'D)^{-1} D' \tilde{Y}^*) = (D'D)^{-1} D' E(\tilde{Y}^*) = (D'D)^{-1} D' E(D\hat{\beta} + \varepsilon^*) = \\ &= (D'D)^{-1} D' (E(D\hat{\beta}) + E(\varepsilon^*)) = (D'D)^{-1} D' D\hat{\beta} = \hat{\beta} \end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\beta}^*) &= \text{Var}(D'D)^{-1}D'Y^* = (D'D)^{-1}D' \text{Var}(Y^*) D(D'D)^{-1} = \\
&= (D'D)^{-1}D' \text{Var}(D\hat{\beta} + \varepsilon^*) D(D'D)^{-1} = \\
&= (D'D)^{-1}D' \text{Var}(\varepsilon^*) D(D'D)^{-1} = (D'D)^{-1}D'D(D'D)^{-1} \hat{\sigma}^2 = \\
&= (D'D)^{-1}D' \hat{\sigma}^2
\end{aligned}$$

2.3 O "BOOTSTRAP" NO CONTEXTO DE SUPERFÍCIE DE RESPOSTAS :

A utilização do *bootstrap* no contexto de superfície de respostas é aqui empregada como uma ferramenta na análise de precisão em modelos de superfície de respostas. A análise da precisão de tais modelos é dificultada pelo fato da estimativa ser uma superfície. Em decorrência disso, o *bootstrap* vem a ser uma técnica útil na criação de imagens desta superfície, proporcionando assim uma visão gráfica da variabilidade da superfície.

Exemplo 2.3.1:

Tomemos como exemplo um conjunto de dados gerados pelo program 2VARIAV2.PAS (Anexo I) que ajusta uma superfície de 2ª ordem utilizando um delineamento Rotacionável de 2ª ordem com n_c pontos centrais. Seja o delineamento definido em $X_1 = (\xi_1 - 150)/10$ e $X_2 = (\xi_2 - 120)/5$ com quatro pontos centrais e sendo os erros provenientes de uma $N(0,1)$. A matriz de delineamento e o vetor de respostas dão dados por:

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -\sqrt{2} & 0 & 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 & 2 & 0 & 0 \\ 1 & 0 & -\sqrt{2} & 0 & 2 & 0 \\ 1 & 0 & \sqrt{2} & 0 & 2 & 0 \end{pmatrix}$$

Ajustando-se por mínimos quadrados o modelo

$$\hat{y} = 99.8878 + 1.0773X_1 + 0.57X_2 - 2.3117X_1^2 - 0.7629X_1X_2 \quad (2.3.1)$$

que produz um ponto de máximo na resposta $y_b = 100.16$ cujas coordenadas são dadas por $X_b = (152.707, 122.217)$. A figura 2.3.1 mostra as curvas de nível para o ajuste. A aplicação do método *bootstrap*, feita iterativamente dentro do mesmo programa 2VARIIV2.PAS, com a recuperação dos erros pelo fator $(n/(n-6))^{1/2}$, forne as imagens *bootstrap* mostradas na figura 2.3.2.

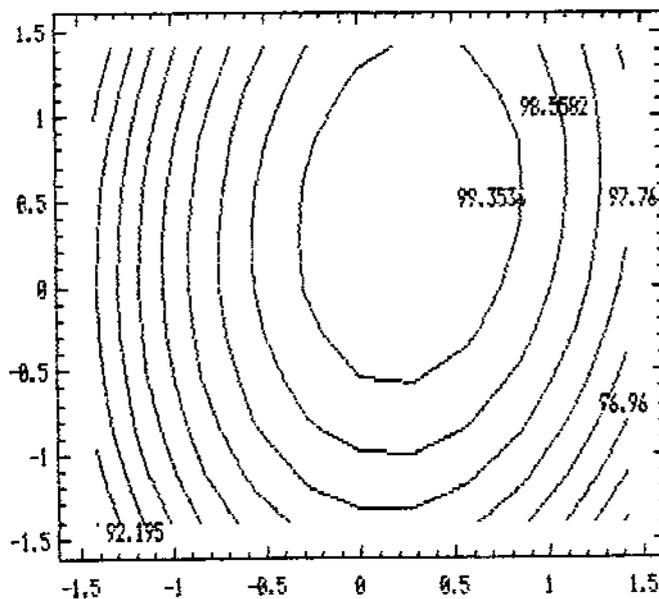


figura 2.3.1 : Curvas de nível para o $\hat{y} = 99.887 + 1.077X_1 + 0.57X_2 - 2.312X_1^2 - 0.7629X_1X_2 + 0.393X_1^2X_2$

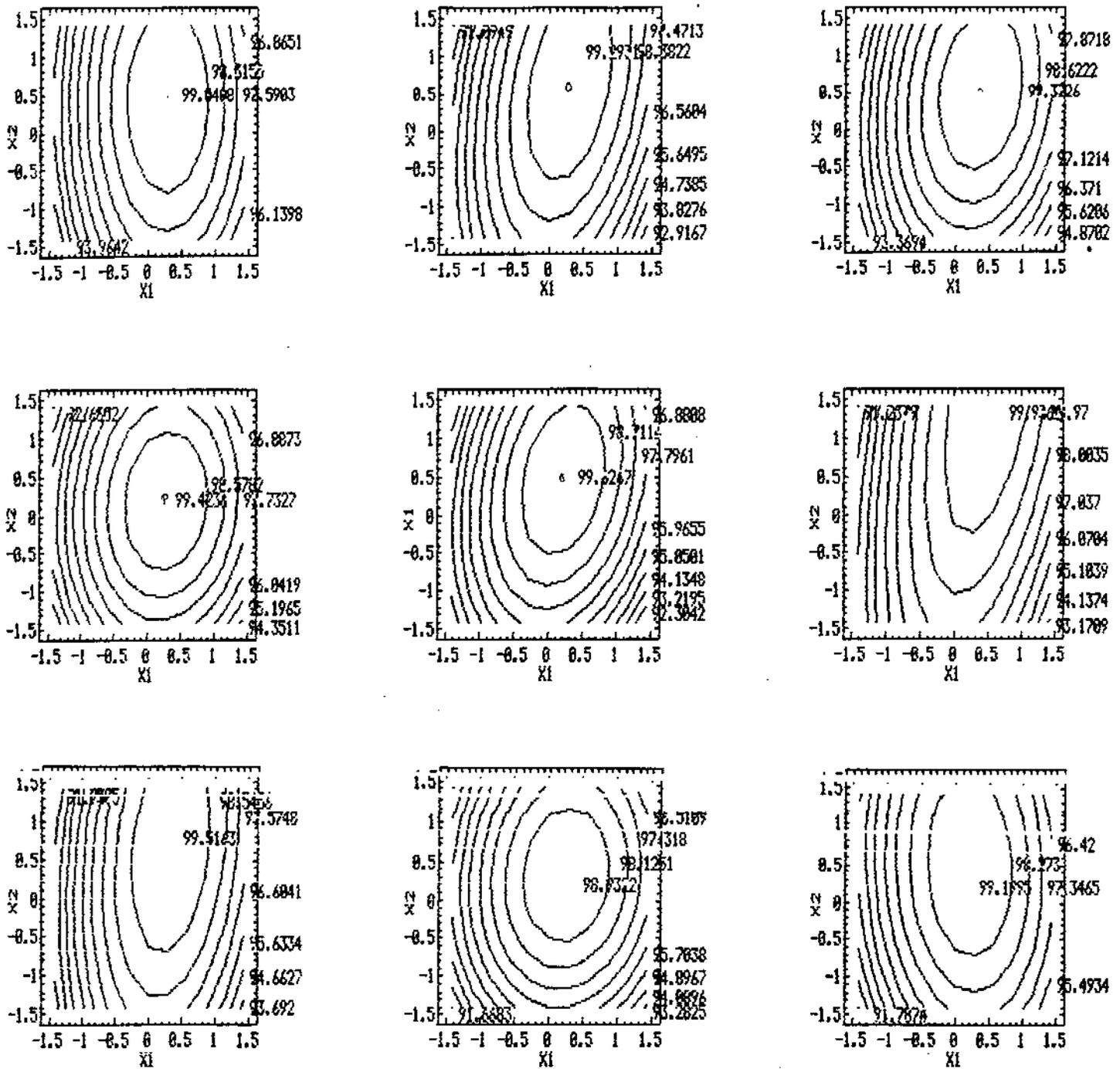


figura 2.3.2 : Imagens bootstrap das curvas de nível do modelo (2.3.1) ajustado no exemplo 2.3.1.

Em termos unicamente dos pontos estacionários, temos na figura 2.3.3 abaixo a posição da estimativa original e de suas imagens *bootstrap*.

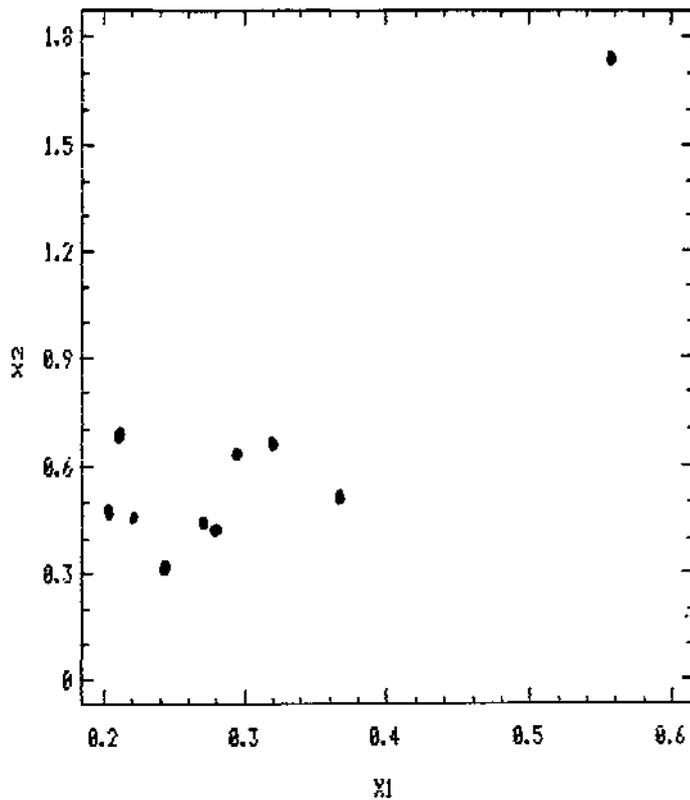


figura 2.3.3 : Estimativa do ponto estacionário e imagens bootstrap do modelo (2.3.1) ajustado no exemplo 2.3.1.

Estas figuras dão uma idéia da variabilidade de nossa estimativa da posição do ponto estacionário da superfície de respostas, X_e .

Em problemas rotineiros de Exploração de Superfície de Respostas, geralmente o objetivo final é a estimação das coordenadas do ponto ótimo de operação, $(X_{1s} , X_{2s} , \dots , X_{ks})$. Nos estágios finais da exploração, quando a região do delineamento já engloba, ou pelo menos se avizinha bastante daquele ponto, modelos polinomiais completos de 2ª ordem são usualmente empregados.

A partir do ajuste do modelo, a estimativa pontual daquele vetor de coordenadas, resolvendo-se o sistema de equações lineares

$$\begin{aligned} \hat{\delta Y} / \delta X_1 &= 0 \\ \hat{\delta Y} / \delta X_2 &= 0 \\ &\vdots \\ \hat{\delta Y} / \delta X_k &= 0 \end{aligned}$$

desde que neste ponto a matriz de derivadas segundas seja negativa definida caso se procure um ponto de máximo - ou positiva definida - caso contrário. A solução é dada por

$$X_s = -B^{-1} \hat{\beta} / 2 , \quad (2.3.2)$$

como mostrada pela equação (1.3.2.6), onde em um modelo com duas variáveis preditoras, $k=2$, esta solução seria dada por:

$$\begin{aligned} X_{1s} &= (\hat{\beta}_{12} \hat{\beta}_2/2 - \hat{\beta}_{22} \hat{\beta}_1) / (2\hat{\beta}_{11} \hat{\beta}_{22} - \hat{\beta}_{12}^2/2) \\ X_{2s} &= (\hat{\beta}_{12} \hat{\beta}_1/2 - \hat{\beta}_{11} \hat{\beta}_2) / (2\hat{\beta}_{11} \hat{\beta}_{22} - \hat{\beta}_{12}^2/2) \end{aligned}$$

A dificuldade aqui consiste em se associar limites de confiança às estimativas. Dada a não linearidade em (2.3.2), mesmo sob hipóteses de normalidade dos resíduos, a teoria normal não é útil neste caso.

Nossa proposta aqui é explorar a possibilidade de utilizar o método *bootstrap* para prover estimativas desta variabilidade. Para tais aplicações o *bootstrap* segue o roteiro abaixo.

(i) Para o modelo de 2ª ordem em k variáveis preditoras, descrito na seção 1.3.2, dado por $\underline{y} = D\hat{\beta}_{(1)} + \underline{\varepsilon}$, onde $\underline{\varepsilon}$ tem distribuição normal com $E(\underline{\varepsilon})=0$ e $\text{Var}(\underline{\varepsilon})=I \sigma_n^2$, i.i.d; o vetor das estimativas dos parâmetros do modelo, $\hat{\beta}_{(1)}$, ajustado por mínimos quadrados é dado por $\hat{\beta}_{(1)} = (D'D)^{-1}D'y$.

(ii) O modelo ajustado é dado por $\hat{y} = D\hat{\beta}_{(1)} = \hat{\beta}_0 + D'\hat{\beta} + D'BD$, como foi definido na seção 1.3.2, com as coordenadas do ponto estacionário dadas por $X_s = -B^{-1}\hat{\beta}/2$, cuja resposta estimada no ponto estacionário é dada por $\hat{y}_s = \hat{\beta}_0 + X_s' \hat{\beta}/2$.

(iii) O vetor dos resíduos, \underline{r} , é determinado como sendo $\underline{r} = \underline{Y} - \hat{\underline{Y}} = \underline{Y} - D\hat{\beta}_{(1)}$. Aplicando-se o fator de correção $(n/(n-p))^{1/2}$ e assumindo-se para \underline{r} uma distribuição empírica, \hat{F} , que associa a probabilidade $1/n$ a cada $r_i(n/(n-p))^{1/2}$, tem-se assim $r_i(n/(n-p))^{1/2} \sim \hat{F}$.

(iv) Tira-se uma amostra aleatória com reposição, chamada amostra *bootstrap*, de \hat{F} denotada por $\underline{r}^* = (r_1^*, \dots, r_n^*)$, e calcula-se um vetor de pseudo-dados $\underline{Y}^* = \hat{\underline{Y}} + \underline{r}^* = D\hat{\beta}_{(1)} + \underline{r}^*$.

(v) Ajusta-se novamente um modelo de 2ª ordem por mínimos quadrados e a estimativa *bootstrap* de $\hat{\beta}_{(1)}$ como sendo $\hat{\beta}_{(1)}^* = (D'D)^{-1}D'Y^*$.

(vi) Calculando-se assim a estimativa *bootstrap* do ponto estacionário, dada por $X_s^* = -B^{*-1}\hat{\beta}^*/2$, sendo a estimativa *bootstrap* da resposta estimada no ponto estacionário $\hat{y}_s^* = \hat{\beta}_0^* + X_s^* \hat{\beta}^*/2$.

Assim, os passos (iv)-(vi) são repetidos um número nbs de vezes fornecendo $X_s^{*1}, X_s^{*2}, \dots, X_s^{*nbs}$ como sendo uma aproximação da distribuição amostral de X_s .

As regiões de confiança foram definidas com base na Distância de Mahalanobis entre as coordenadas do ponto estacionário real X_s e as estimativas *bootstrap* $X_s^{*1}, X_s^{*2}, \dots, X_s^{*nbs}$, dadas por:

$$DMH(X_s) = (X_s - \bar{X}_s^*)' \hat{\Sigma}_{bs}^{-1} (X_s - \bar{X}_s^*)$$

$$DMH(X_s^{*i}) = (X_s^{*i} - \bar{X}_s^*)' \hat{\Sigma}_{bs}^{-1} (X_s^{*i} - \bar{X}_s^*)$$

onde: X_s : é o ponto estacionário verdadeiro.

X_s^{*i} : é a estimativa *bootstrap* do ponto estacionário.

\bar{X}_s^* : o vetor de média das estimativas *bootstrap* do ponto estacionário.

$\hat{\Sigma}_{bs}^{-1}$: a matriz de variâncias e covariâncias das estimativas *bootstrap* do ponto estacionário.

Assim, se o número de vezes que $DMH(X_s^{*i}) > DMH(X_s)$ for superior a $\alpha*100\%$ do total de *nbs*, então é porque a região cobre o ponto estacionário em $(1-\alpha)*100\%$ das vezes.

CAPÍTULO 3 - AS SIMULAÇÕES EM MODELOS DE 2ª ORDEM

3.1 A SIMULAÇÃO EM UM MODELO DE 2ª. ORDEM COM DOIS PREDITORES :

As simulações foram feitas com base em modelos polinomiais de 2ª ordem para delineamentos compostos centrais (Delineamentos Rotacionáveis de 2ª ordem - Cubo + Estrela + Pontos Centrais) em duas variáveis preditoras.

Os modelos definidos em (1.5.3) dados por

$$y = y_s + (X - X_s)' M \Delta M' (X - X_s) + \varepsilon$$

foram utilizados para as simulações, onde $\varepsilon \sim N(0, \sigma^2)$, $\sigma = 1, 0.8$ e 0.5 , com $M = [m_1 \ m_2]$ sendo $m_1' = [\cos 30^\circ \ \sin 30^\circ]$ e $m_2' = [-\sin 30^\circ \ \cos 30^\circ]$ e $\Delta = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ os valores de y_s e X_s

verdadeiros foram escolhidos e a partir disso um vetor de pseudo-dados \underline{Y} determinado, ajustando-se novamente um polinômio de 2ª ordem, via mínimos quadrados, e calculando-se a estimativa do ponto estacionário conforme (1.3.2.7). Com base neste modelo estimado o método *bootstrap* foi aplicado seguindo-se o roteiro da seção 2.3.

As regiões de confiança foram definidas com base na Distância de Mahalanobis entre as coordenadas do ponto estacionário verdadeiro X_0 e as estimativas *bootstrap* $X_0^{*1}, X_0^{*2}, \dots, X_0^{*BS}$.

Nas simulações feitas pelo programa 2VARIIV2.PAS (Anexo I) conforme descrição acima, com base em 1000 simulações Monte Carlo e 1000 repetições *bootstrap*, fixamos os valores do ponto ótimo verdadeiro em $y_0 = 100$ com coordenadas dadas por $X_{10} = 0.2$ e $X_{20} = 0.4$, variando os autovalores para a obtenção de superfícies mais íngremes ou menos íngremes, e regiões mais ou menos simétricas. O delineamento foi centrado em função das variáveis preditoras naturais em $(\xi_{1c}, \xi_{2c}) = (150, 120)$ com $\Delta\xi_1 = 10$ e $\Delta\xi_2 = 5$, fazendo com que o ponto ótimo verdadeiro estivesse contido na região do delineamento. Os resultados obtidos constam nas tabelas abaixo.

TABELA 3.1 : Resultados obtidos pela simulação de $b_s=1000$ e 1000 Monte Carlo com erros provenientes de $N(0,1)$.

λ_1	λ_2	(*)	(**)
-0.025	-0.017	98.80%	98.20%
-0.25	-0.17	99.20%	97.50%
-0.64	-0.59	98.80%	95.90%
-1.25	-1.17	95.10%	91.00%

(*) : Percentual das vezes em que a região de 95% de confiança cobriu o ponto ótimo verdadeiro.

(**): Percentual das vezes em que a região de 90% de confiança cobriu o ponto ótimo verdadeiro.

TABELA 9.2 : Resultados obtidos pela simulação de $b_2=1000$ e 1000 Monte Carlo com erros provenientes de $N(0,0.8^2)$.

λ_1	λ_2	(*)	(**)
-0.025	-0.017	98.80%	97.30%
-0.25	-0.17	99.20%	98.20%
-0.64	-0.59	96.40%	94.30%
-1.25	-1.17	92.60%	88.20%

(*) : Percentual das vezes em que a região de 95% de confiança cobriu o ponto ótimo verdadeiro.

(**): Percentual das vezes em que a região de 90% de confiança cobriu o ponto ótimo verdadeiro.

TABELA 9.3 : Resultados obtidos pela simulação de $b_s=1000$ e 1000 Monte Carlo com erros provenientes de $N(0,0.5^2)$.

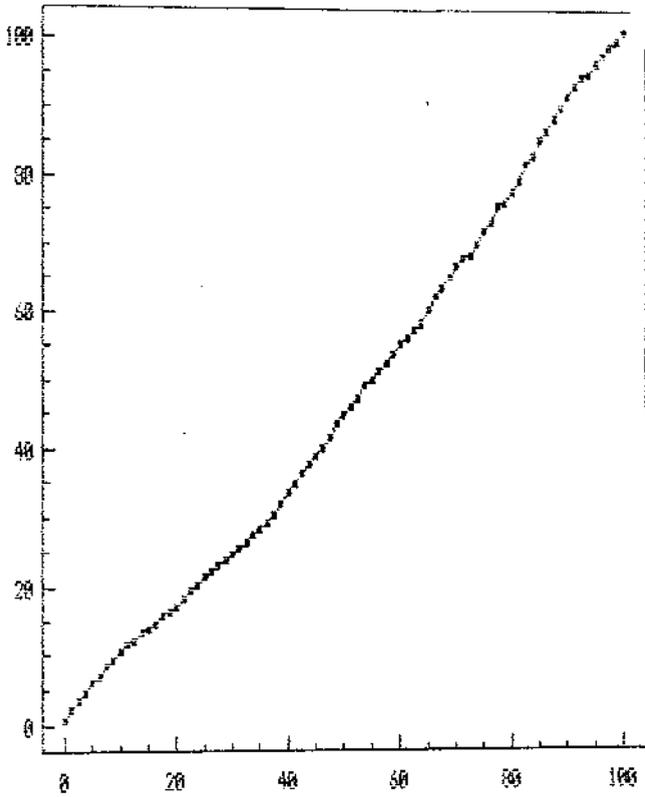
λ_1	λ_2	(*)	(**)
-0.025	-0.017	98.80%	97.30%
-0.25	-0.17	97.70%	96.00%
-0.64	-0.59	93.70%	90.40%
-1.25	-1.17	90.30%	84.70%

(*) : Percentual das vezes em que a região de 95% de confiança cobriu o ponto ótimo verdadeiro.

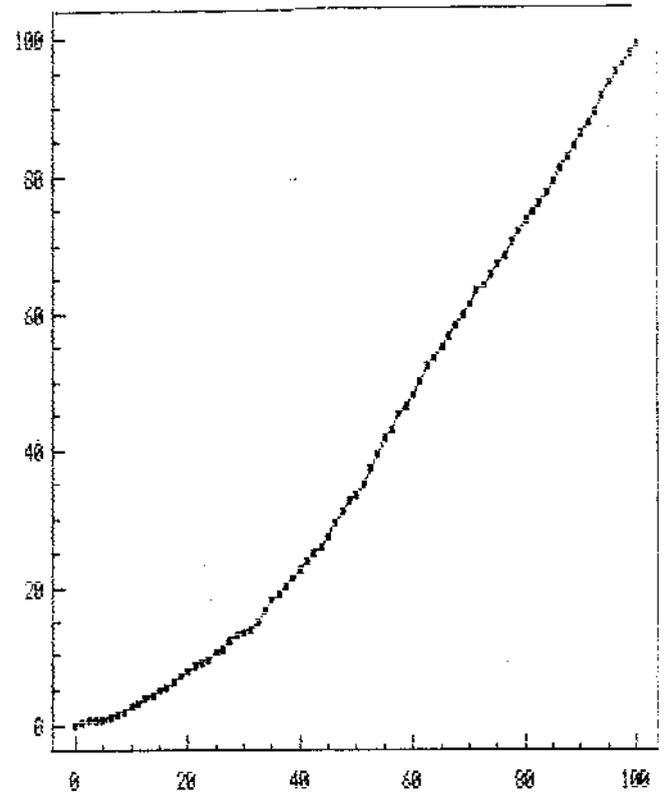
(**): Percentual das vezes em que a região de 90% de confiança cobriu o ponto ótimo verdadeiro.

Nos gráficos a seguir mostramos o nível de significância real - estimado através de 1000 simulações Monte Carlo - versus o nível de significância nominal da diferença entre os vetores X_s e \hat{X}_s , para diversas hipóteses sobre λ_1 , λ_2 e σ . Em todos os casos, para cada simulação Monte Carlo foram efetuadas 1000 repetições *bootstrap*.

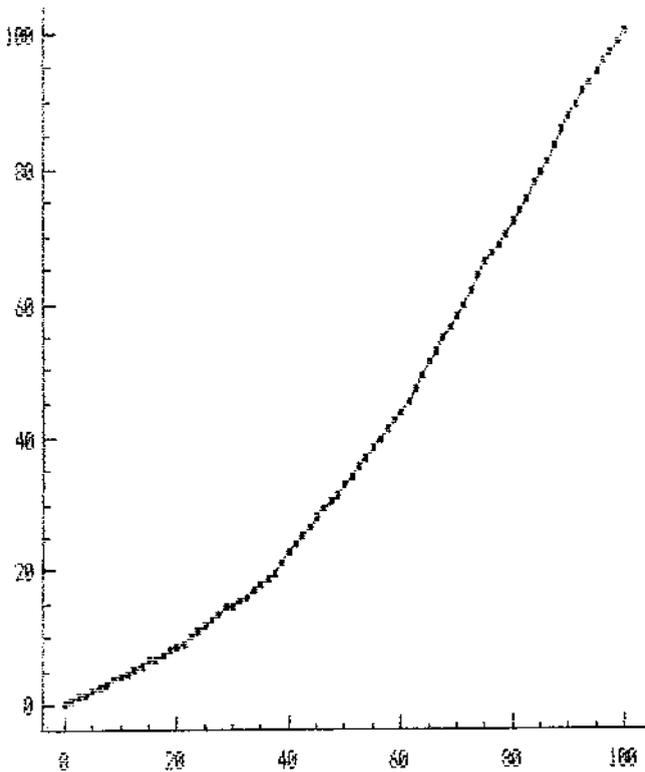
$\mu_1 = -0.625$ e $\mu_2 = -0.617$, $\sigma = 1$



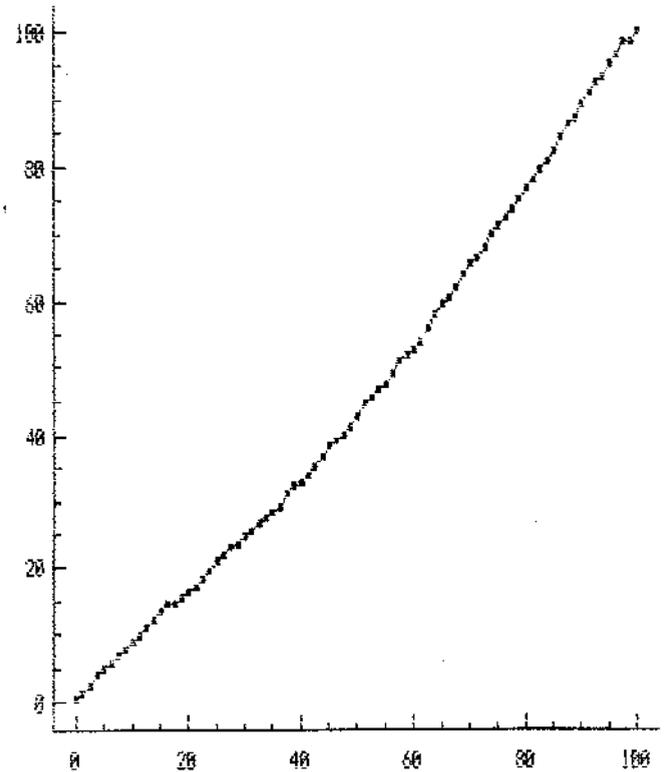
$\mu_1 = -0.25$ e $\mu_2 = -0.17$, $\sigma = 1$



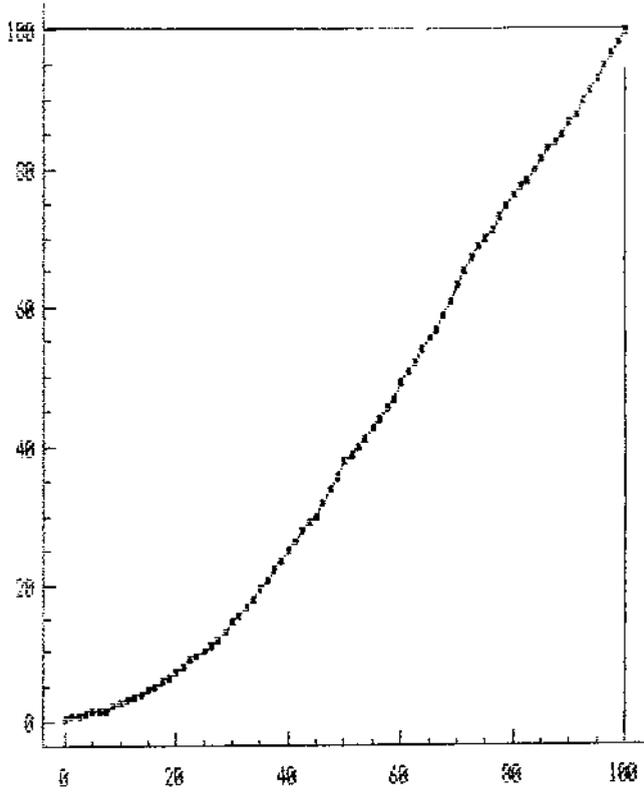
$\mu_1 = -0.64$ e $\mu_2 = -0.55$, $\sigma = 1$



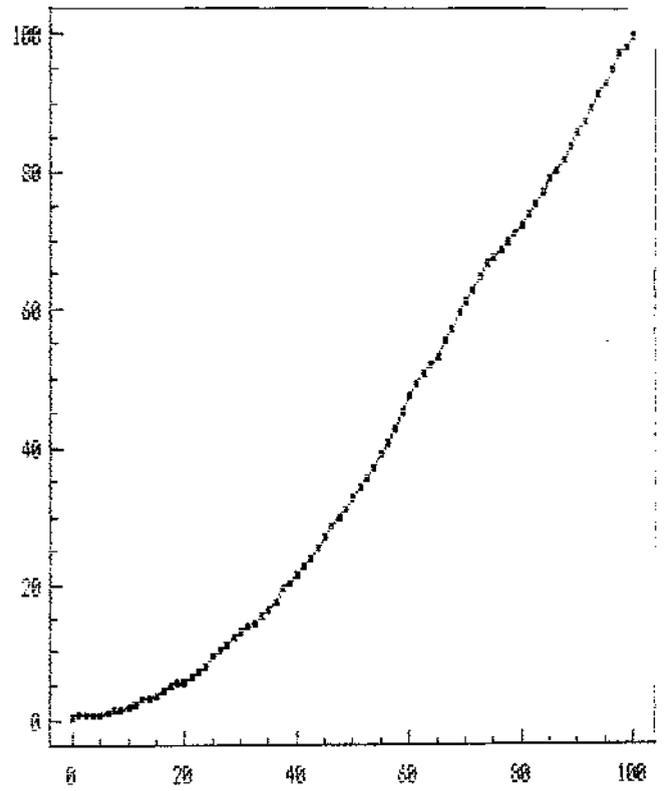
$\mu_1 = -1.25$ e $\mu_2 = -1.17$, $\sigma = 1$



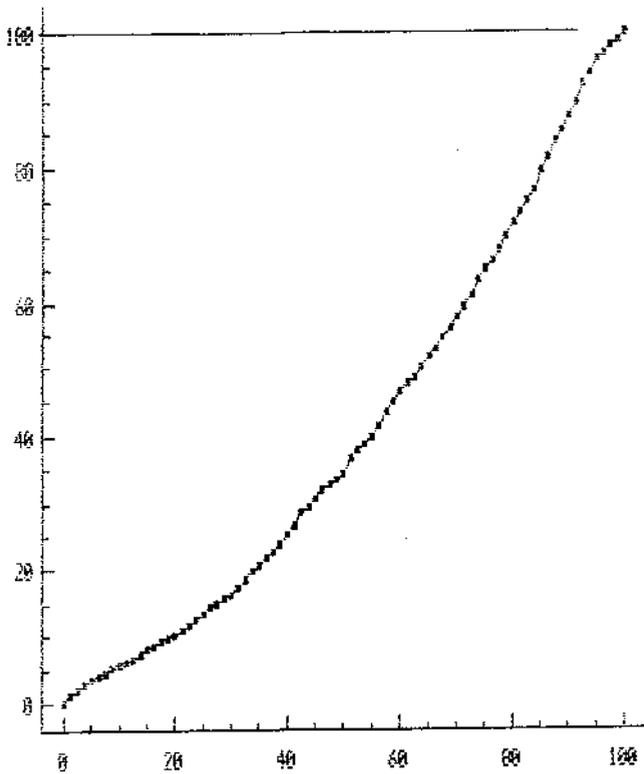
$\mu_1 = -0.025$ e $\mu_2 = -0.017$, $\sigma = 0.8$



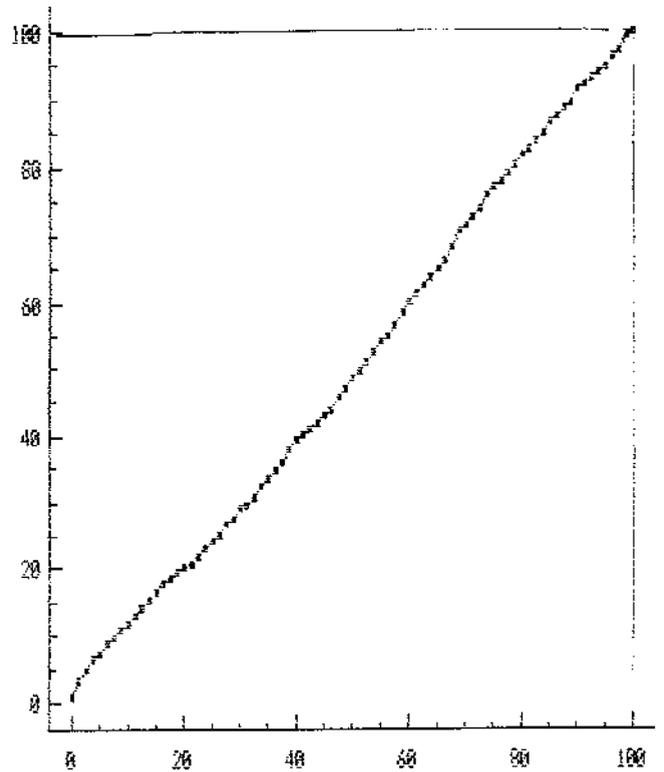
$\mu_1 = -0.25$ e $\mu_2 = -0.17$, $\sigma = 0.8$



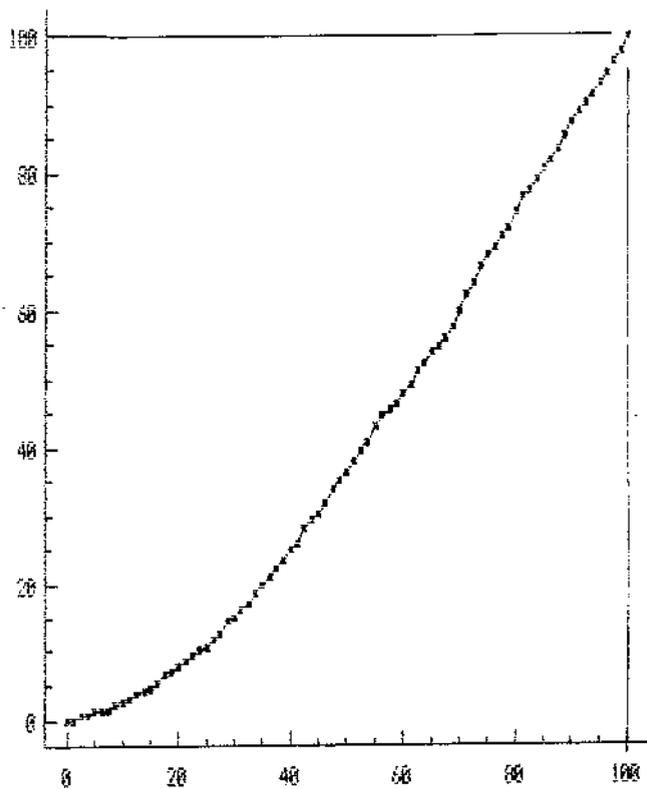
$\mu_1 = -0.64$ e $\mu_2 = -0.59$, $\sigma = 0.8$



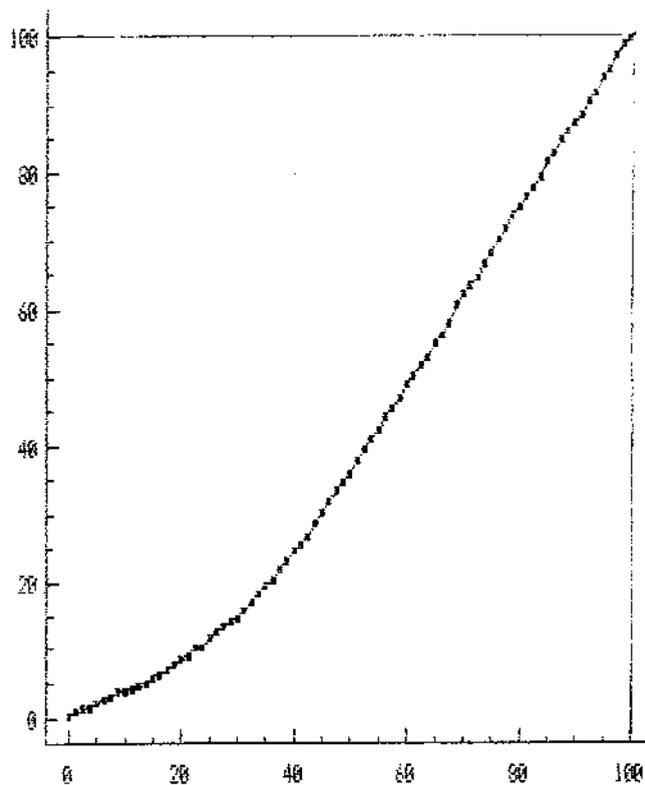
$\mu_1 = -1.25$ e $\mu_2 = -1.17$, $\sigma = 0.8$



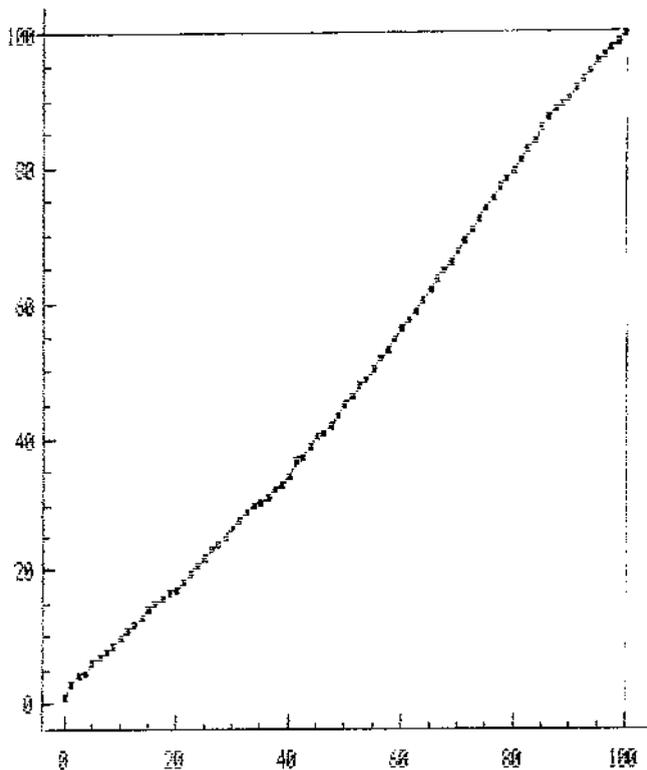
$\mu_1 = -0.625$ e $\mu_2 = -0.817$, $\sigma = 0.5$



$\mu_1 = -0.25$ e $\mu_2 = -0.17$, $\sigma = 0.5$



$\mu_1 = -0.69$ e $\mu_2 = -0.59$, $\sigma = 0.5$



$\mu_1 = -1.25$ e $\mu_2 = -1.17$, $\sigma = 0.5$

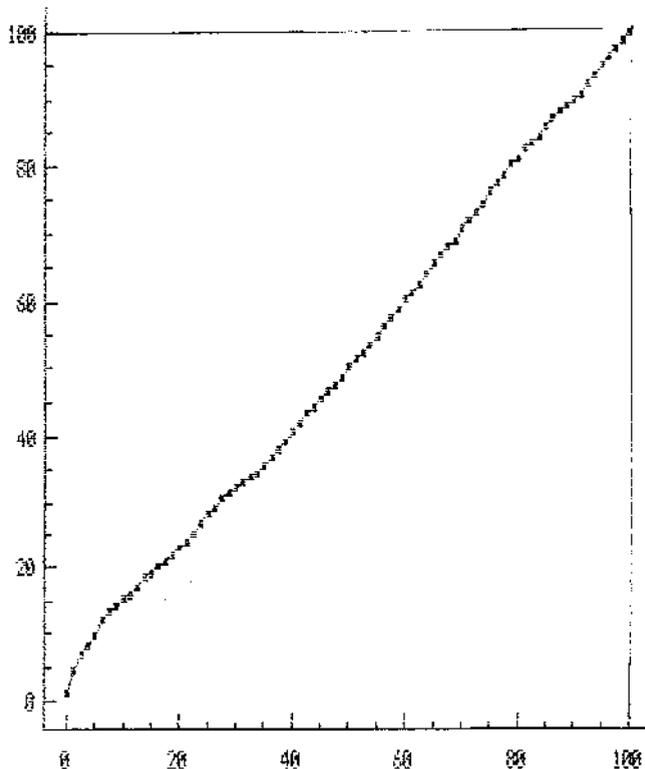


TABELA 3.4 : Resultados obtidos pela simulação de $nbs=1000$ e 1000 Monte Carlo para três preditores, com erros provenientes de $N(0,1)$.

λ_1	λ_2	λ_3	(*)	(**)
-0.025	-0.017	-0.013	97.60%	94.44%
-0.64	-0.59	-0.57	96.30%	92.30%
-1.25	-1.17	-1.13	95.10%	91.00%

TABELA 3.5 : Resultados obtidos pela simulação de $nbs=1000$ e 1000 Monte Carlo para três preditores, com erros provenientes de $N(0,0.9^2)$.

λ_1	λ_2	λ_3	(*)	(**)
-0.025	-0.017	-0.013	97.90%	94.40%
-0.64	-0.59	-0.57	97.00%	93.40%
-1.25	-1.17	-1.13	93.70%	88.50%

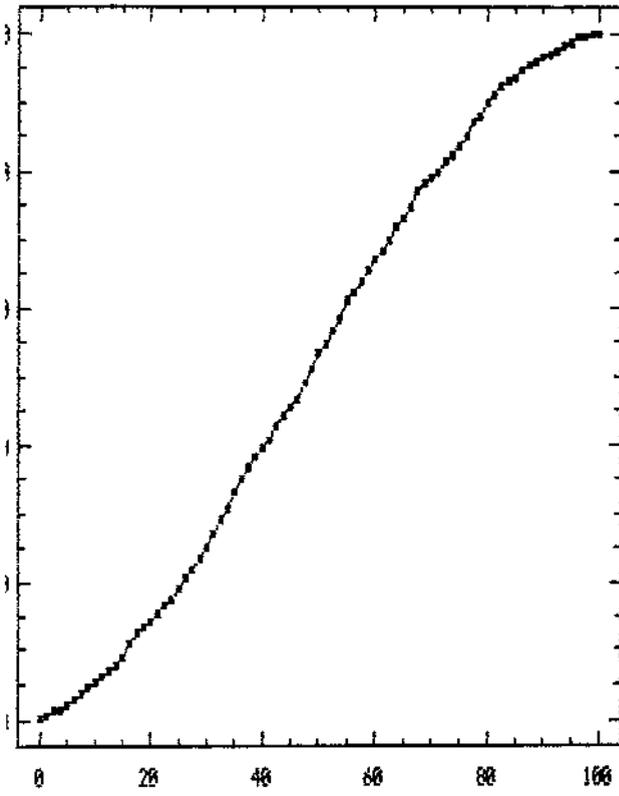
TABELA 3.6 : Resultados obtidos pela simulação de $nbs=1000$ e 1000 Monte Carlo para três preditores, com erros provenientes de $N(0,0.5^2)$.

λ_1	λ_2	λ_3	(*)	(**)
-0.025	-0.017	-0.013	98.50%	95.00%
-0.64	-0.59	-0.57	96.00%	92.10%
-1.25	-1.17	-1.13	93.30%	88.70%

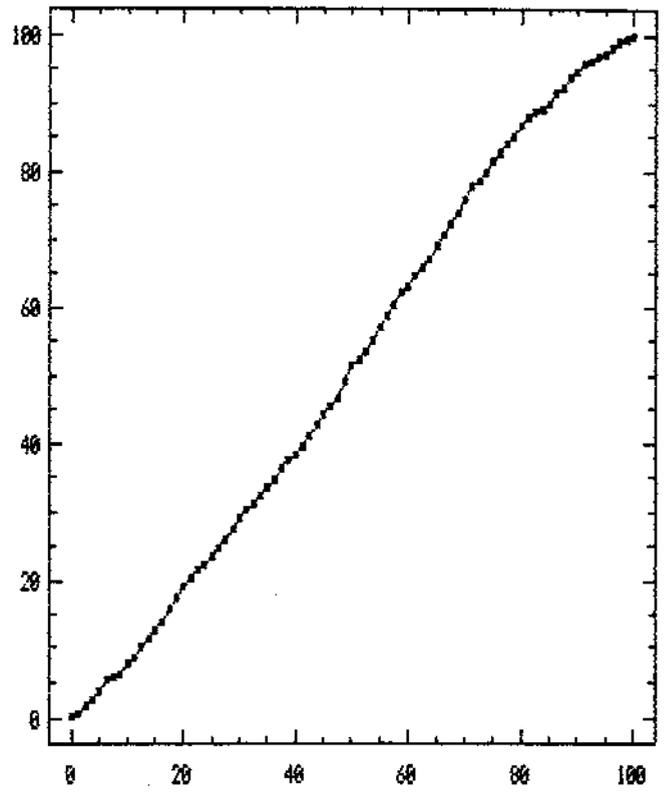
(*) : Percentual das vezes em que a região de 95% de confiança cobriu o ponto ótimo verdadeiro.

(**) : Percentual das vezes em que a região de 90% de confiança cobriu o ponto ótimo verdadeiro.

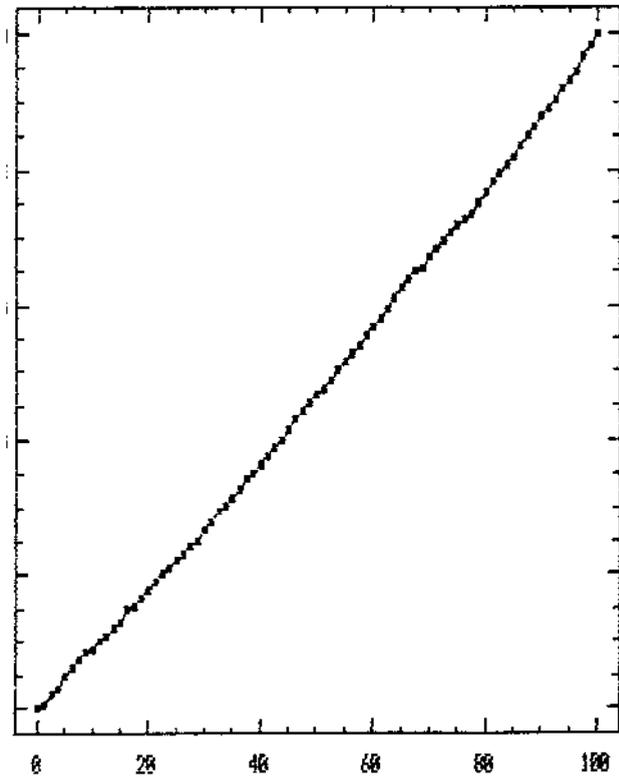
$\lambda_1 = -0.625, \lambda_2 = -0.817$ e $\lambda_3 = -0.013, \sigma = 1$



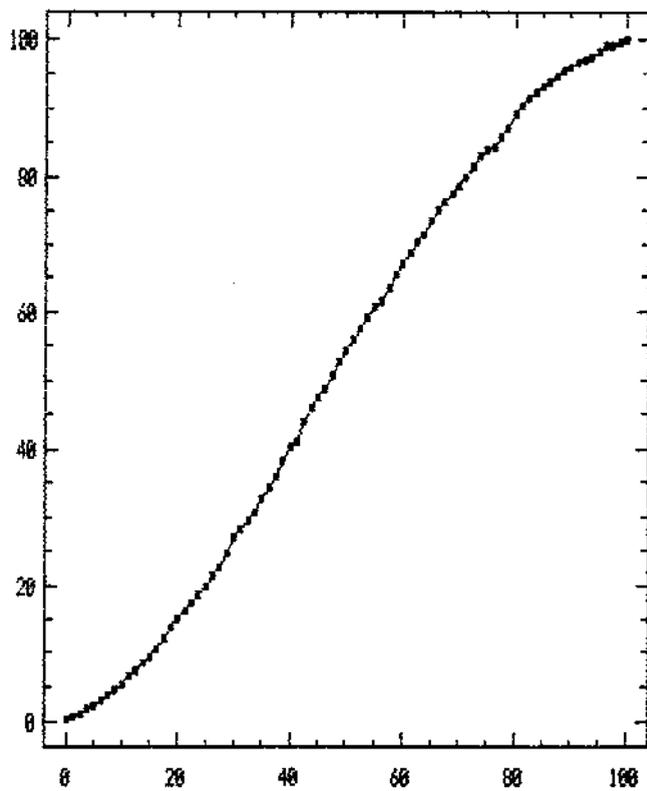
$\lambda_1 = -0.64, \lambda_2 = -0.59$ e $\lambda_3 = -0.57, \sigma = 1$



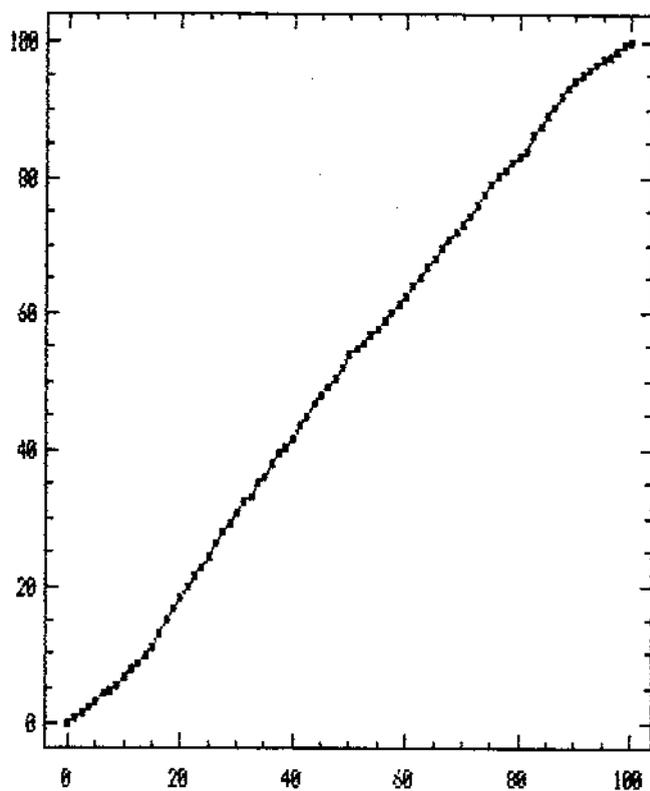
$\lambda_1 = -1.25, \lambda_2 = -1.17$ e $\lambda_3 = -1.13, \sigma = 1$



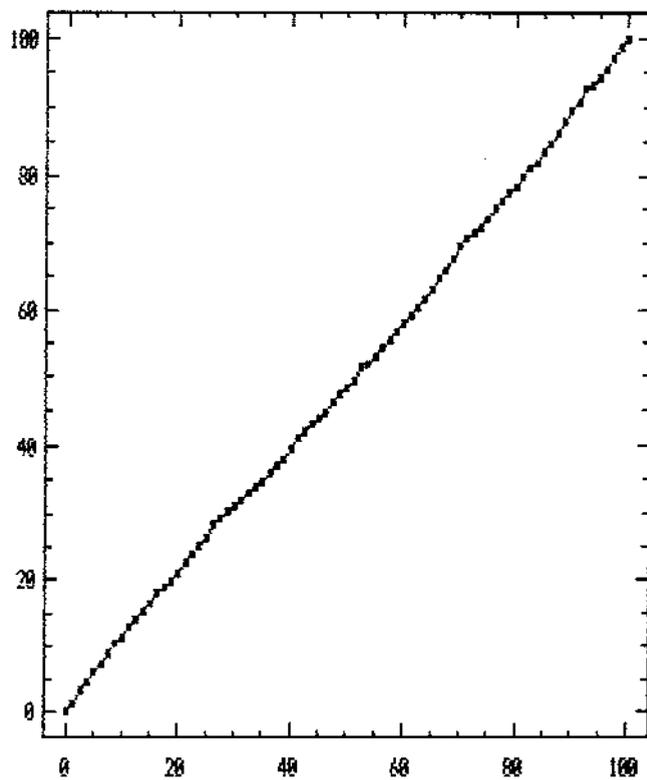
$\mu_1 = 0.025, \mu_2 = -0.017$ e $\mu_3 = -0.013, \sigma = 0.8$



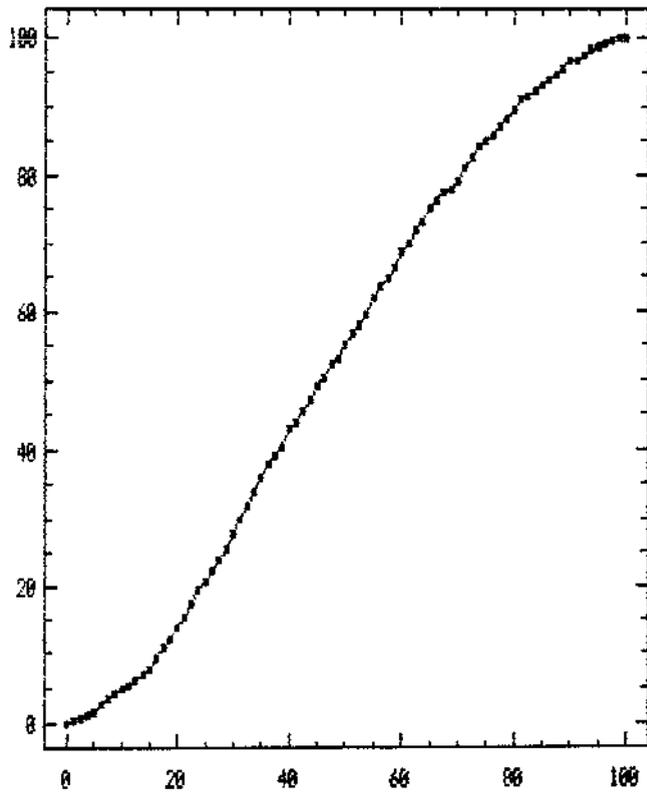
$\mu_1 = -0.64, \mu_2 = -0.59$ e $\mu_3 = -0.013, \sigma = 0.8$



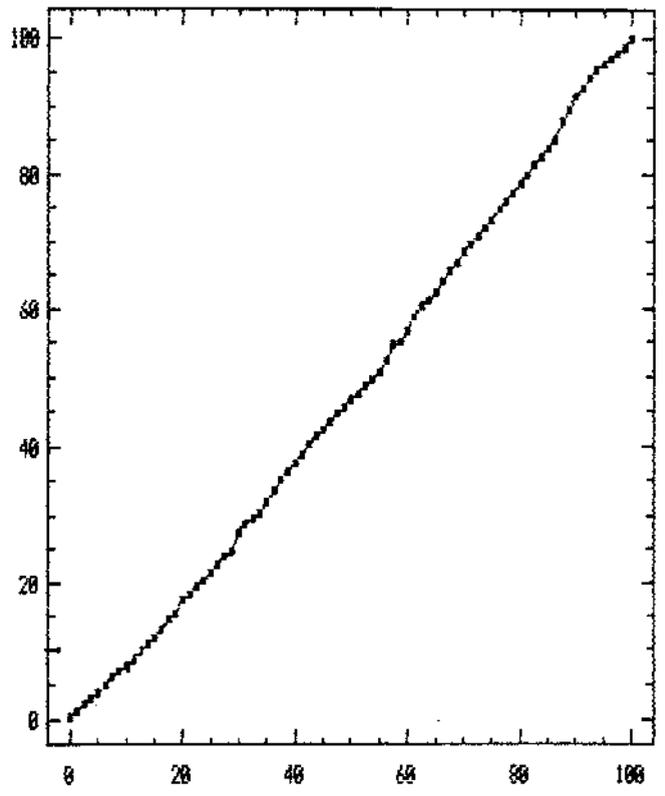
$\mu_1 = -1.25, \mu_2 = -1.17$ e $\mu_3 = -1.13, \sigma = 0.8$



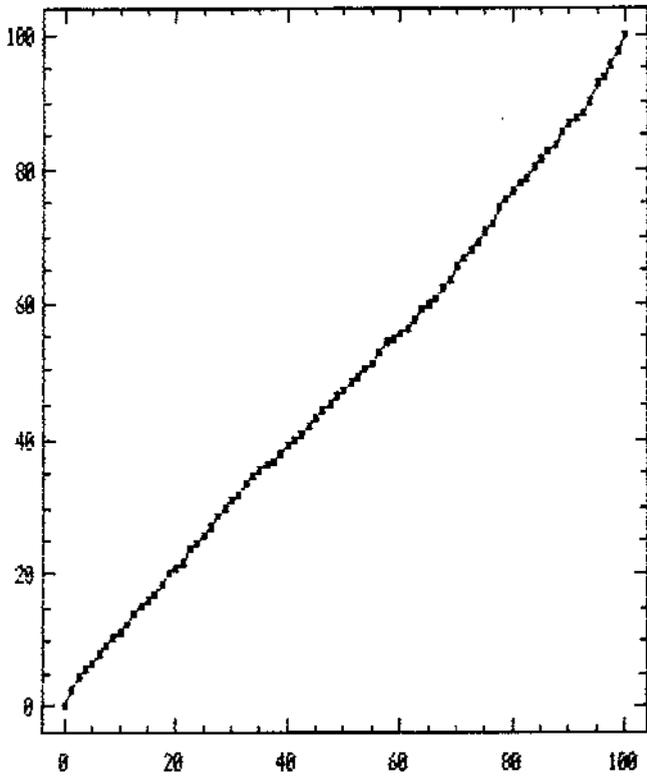
$\mu_1 = -0.025$, $\mu_2 = -0.017$ e $\mu_3 = -0.013$, $\sigma = 0.5$



$\mu_1 = -0.64$, $\mu_2 = -0.59$ e $\mu_3 = -0.57$, $\sigma = 0.5$



$\mu_1 = -1.25$, $\mu_2 = -1.17$ e $\mu_3 = -1.13$, $\sigma = 0.5$



Vemos que em alguns casos a coincidência entre os níveis de significância real e nominal é muito boa em toda a faixa. É o caso $\lambda_1 = -1.25$, $\lambda_2 = -1.17$ e $\sigma = 0.8$ para dois preditores, e para o caso $\lambda_1 = -1.25$, $\lambda_2 = -1.17$, $\lambda_3 = -1.13$ e $\sigma = 1$, entre outros.

De um modo geral verifica-se o seguinte:

- A concordância entre os níveis nominal e real é muito boa quando a superfície é bem definida (auto valores grandes relativos a σ) na vizinhança do ponto estacionário.
- O nível de significância real tende a ser menor que o nominal na faixa crítica, em torno de 5%, quando a superfície é achatada (auto valores pequenos comparados a σ) na vizinhança do ponto estacionário. Esta condição gera então regiões de confiança mais amplas, com níveis de confiança reais superiores aos nominais.
- Em todos os casos o procedimento proposto gera regiões de confiança tecnicamente satisfatórias para a localização exata do ponto estacionário.

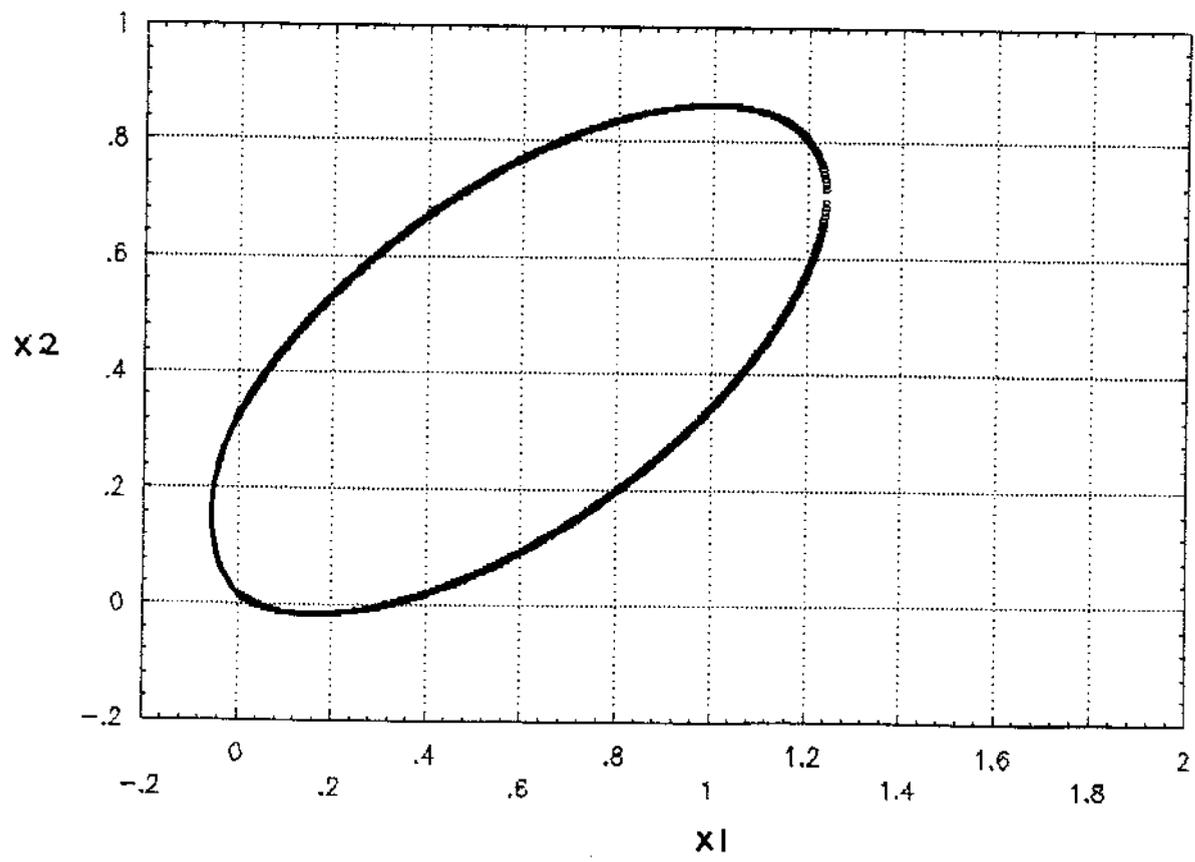
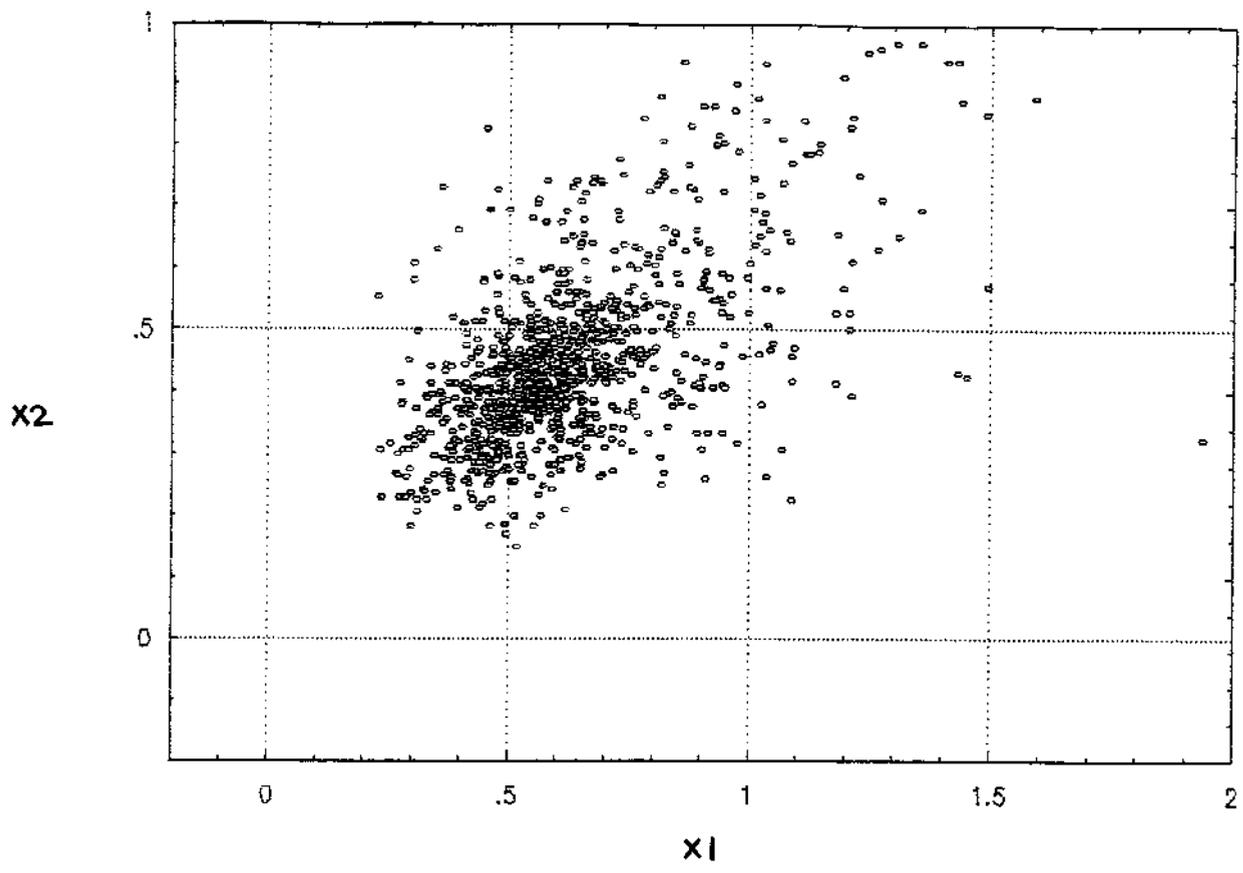
Damos agora um exemplo do emprego deste procedimento para a construção, a partir dos dados experimentais em um delineamento rotacionável de 2ª ordem, com duas variáveis preditoras, de uma região de 95% de confiança para a localização exata do ponto estacionário

(X_{1s} , X_{2s}) de uma Superfície de Respostas.

Para construção da região de confiança:

- 1 - Ajusta-se o modelo.
- 2 - Estima-se a posição do ponto estacionário: \hat{X}_s .
- 3 - Calcula-se os resíduos.
- 4 - Aplica-se o fator de correção dos resíduos.
- 5 - Procede-se a reamostragem *bootstrap*, definindo BS estimativas *bootstrap* da localização do ponto estacionário.
- 6 - Determina-se a matriz $\hat{\Sigma}_{bs}$.
- 7 - Calcula-se a distância de Mahalanobis de \hat{X}_s a cada uma das estimativas \hat{X}_s^* .
- 8 - Ordena-se as distâncias.
- 9 - Determina-se o 100(1- α)-ésimo percentual das distâncias.
- 10- Determina-se o lugar geométrico dos pontos do espaço das variáveis preditoras que estão a uma distância de Mahalanobis igual ou inferior àquele percentil, de \hat{X}_s .

A região de confiança de 100(1- α)% de confiança para X_s é o lugar geométrico assim definido. A figura abaixo ilustra uma situação de $nmc=1$ e $bs=1000$ para um modelo definido no exemplo 2.3.1 com $\lambda_1= -0.025$ e $\lambda_2= -0.017$, delimitando a região de confiança.



Vemos que as distorções nos níveis de confiança reais das regiões de confiança são desprezíveis quando os autovalores têm valores próximos de σ ; nestes casos nenhuma correção é necessária. Quando os autovalores têm valor absoluto muito inferior ou superior a σ , as distorções se acentuam. No primeiro caso as distorções são no sentido de níveis de confiança reais maiores que os nominais; regiões construídas são portanto mais amplas que o necessário. No segundo caso as distorções se invertem, com níveis de confiança reais abaixo dos nominais e, portanto, regiões mais estreitas do que o necessário.

Esta regularidade sugere correções que podem ser feitas, numa situação de experimento real. A partir dos resíduos pode-se estimar σ ; e a partir dos valores *bootstrap* obtém-se $\hat{\Sigma}_{ba}$ e daí seus autovalores. A comparação entre estes e a estimativa de σ sugerirá a correção. Caso $\hat{\sigma}$ seja muito maior do que os valores absolutos dos λ 's, podemos construir uma região de confiança com nível inferior ao desejado, para que o nível de significância real caia próximo do desejado. Por exemplo, desejando-se um nível de confiança de 95%, pode-se construir uma região com nível de confiança nominal igual a 92%. Faz-se o contrário na situação oposta, quando $\hat{\sigma}$ é muito pequeno comparado aos valores absolutos dos λ 's.

3.2 CONCLUSÕES :

Mostramos, através de simulações em modelos de dois preditores que a abordagem por *bootstrap* que propusemos tem performance satisfatória para a construção de regiões de confiança para a localização do ponto ótimo.

Os testes foram feitos em diversos contextos com respeito a forma da superfície na vizinhança do ponto ótimo, e também diversos valores para σ . As formas da superfície foram estabelecidas através da escolha apropriada dos autovalores λ_1 e λ_2 .

Com base nesta escolha dos autovalores, compusemos vizinhanças mais íngremes ou menos íngremes, e mais simétricas ou menos simétricas. O *bootstrap* desempenhou bem em todas as situações consideradas.

Há evidências de que a distribuição das estimativas *bootstrap* das coordenadas do ponto ótimo vai se tornando mais assimétrica, à medida em que cresce o valor de σ em relação à amplitude da Superfície de Respostas na região experimental. Esta condição pode ser contornada ou aumentando a amplitude ou reduzindo o σ . Aumenta-se a amplitude alargando-se a região experimental; reduz-se σ realizando-se replicações. De qualquer forma, um valor de σ grande relativamente a amplitude, indica uma baixa relação serial/ruído, indicando baixa detectabilidade das variações na Superfície de Respostas, o que torna questionável o próprio prosseguimento da exploração, sem o recurso das replicações.

O *bootstrap* oferece, assim, uma ferramenta útil na construção de regiões de confiança baseados no critério da distância de Mahalanobis, para localização do ponto de resposta ótimo em Exploração de Superfície de Respostas.

ANEXO I - Programa 2VARIIV2.PAS.

```

( 2variav2.pas )
program duasvariaveis;
($n+,e+)
uses crt,graph;
type
  vetor1=array [1..12] of real;
  vetor2=array [1..6,1..12] of real;
  vetor3=array [1..6] of real;
  vetor4=array [1..1000] of REAL;

const
  esc = #27;

var
  tecla, v : char;

  n, i, j, nbs, k, l, indice, ibs, x10_bs, x10est_coord,
  x20est_coord, x20_bs, gd, gm, imc, contbs, CONTBSCOD,
  nmc, xicentral, x2central, deltax1, deltax2, escx,
  escy, nccentro, nlcentro, nc, nl, delta : integer;
  S1X1BS, S2X1BS, S1X2BS, S2X2BS, SX1X2BS, SOMA1X1BS, SOMA2X1BS,
  MEDX1BS, VX1BS, CVX1X2BS, MEDX2BS, SOMA1X2BS, SOMA2X2BS,
  VX2BS, MAH, MAHBS,
  cst, cst1, cst2, u1, u2, u3,
  MEDIAX1BS, MEDIAX2BS, SOMAX1X2BS, DMAHBS,
  VARX1BS, VARX2BS, COVX1X2BS, DMAH : extended;

  a, b, c, y0, l1, l2, a0, a1, angulo, LimI, LimS, PERCENTUALBS,
  x1s,x2s,
  a2, a11, a22, a12, sigma, e, Ysest, X10S, X20S, X10BS, X20BS, ang,
  Yshs, x1est, x2est, SQRESbs, SQREG, SOMAY,
  x10est, x20est, delta1, delta2, SQRES : real;

  x1, x2, ey, y, ych, erro, erroamos, x1delin, x2delin,
  ychbs, errobs : vetor1;
  m : vetor2;
  alfa : vetor3;
  dados, matriz, refugo, refugo2, refugo3, refugo4 : text;

procedure GERA_ERROS;
begin
  u1:= random;
  u2:= random;
  e:= sqrt(-2*ln(u1)) * sin(2*pi*u2);
end;

( parametros estimados com as var x1 x2 codificadas )
procedure ESTIMA_PARAMETROS;
begin

```

```

    for k:=1 to 6 do
      begin
        alfakl:= 0;
        for l:=1 to n do
          begin
            alfakl:= alfa [k] + m[k,l] * yll;
          end;
        end;
      end;

procedure AMOSTRA_ERROS;
begin
  for i:= 1 to n do
    begin
      u3:= random ;
      indice:= trunc(12*u3) + 1;
      erroamosfil:= erro[indice]*sqrt(n/(n-6));
      yil:= ych[i] + erroamosfil;
    end;
  end;

( PROGRAMA PRINCIPAL )

begin
  assign (dados, 'a:\teste2.dat');
  assign (matriz, 'a:\matriz.dat');
  assign (refugo4, 'a:\X1IX21.DAT');
  rewrite (refugo4);
  reset (dados);
  reset (matriz);
  randomize;
  clrscr;writeln;
  write ('n= ');readln (n);
  write ('sigma = ');readln (sigma);
( y0:= 100;
  x1s:= 152; x2s:= 122;
  xicentral:= 150;
  x2central:= 120;
  ANG:= 30;
  L1:= -0.017; L2:= -0.025;
)
  write ('VALOR DA RESPOSTA NO PT. ESTACIONARIO DESEJADO: ');
  readln (y0);
  write ('VALOR DE X1 NO PT. ESTACIONARIO DESEJADO: ');
  readln (x1s);
  write ('VALOR DE X2 NO PT. ESTACIONARIO DESEJADO: ');
  readln (x2s);
  write ('ANGULO DA ROTAAO: ');
  readln (ang);
  write ('AUTOVALOR1 11: ');
  readln (l1);
  write ('AUTOVALOR2 12: ');
  readln (l2);

```

```

write ('nbs= ');readln (nbs);
write ('CENTRO DO DELINEAMENTO EM X1: ');
readln (x1central);
write ('CENTRO DO DELINEAMENTO EM X2: ');
readln (x2central);
write ('deltax1= ');readln (deltax1);
write ('deltax2= ');readln (deltax2);
( nccentro:= 320;
nlcentro:= 175; )
angulo:=ang*3.141592654/180;

( leitura de x1 e x2 codificadas )
for j:=1 to n do
begin
readln (dados, x1delin[j], x2delin[j]);
end;

( leitura de inversa(x'x)x' com x1 e x2 codificados )
for j:=1 to 6 do
begin
for i:=1 to n do
begin
read (matriz,m[j,i]);
end;
end;

( para as variaveis naturais )

a:=l1 * cos(angulo) * cos(angulo)+l2* sin(angulo)* sin(angulo);
b:=sin(angulo)* cos(angulo)* (l1-l2);
c:=l1* sin(angulo) *sin(angulo)+l2* cos(angulo)* cos(angulo);

a0:= y0 + a*x1s*x1s + 2*b*x1s*x2s + c*x2s*x2s;
a1:= -2*a*x1s - 2*b*x2s;
a2:= -2*b*x1s - 2*c*x2s;
a11:= a;
a22:= c;
a12:= 2*b;
writeln ('a0= ',a0:8:4);writeln ('a1= ',a1:8:4);writeln ('a2= ',a2:8:4);
writeln ('a11= ',a11:8:4);writeln ('a22= ',a22:8:4);
writeln ('a12= ',a12:8:4);

tecla:= readkey;
clrscr;

somay:= 0;
for i:=1 to 12 do
begin
x1[i]:= x1delin[i]*deltax1+x1central;
x2[i]:= x2delin[i]*deltax2+x2central;
ey[i]:= a0 + a1*x1[i] + a2*x2[i] + a11*x1[i]*x1[i] +
a22*x2[i]*x2[i] + a12*x1[i]*x2[i];
GERA_ERROS;

```

```

    yfil:= eyfil + e*sigma;
    somay:= somay+yfil;
end;

ESTIMA_PARAMETROS;
( writeln (lst);

for i:= 1 to 6 do
    begin
        writeln (lst,'alfa',i:2,']= ',alfafil:10:4);
    end;
    writeln(lst) ; writeln ;
)

x1est:= ( alfa[3]*alfa[6] - 2*alfa[2]*alfa[5] )
        / ( 4*alfa[4]*alfa[5] - alfa[6]*alfa[6] );
x2est:= ( alfa[2]*alfa[6] - 2*alfa[3]*alfa[4] )
        / ( 4*alfa[4]*alfa[5] - alfa[6]*alfa[6] );

Ysest:= alfa[1]+alfa[2]*x1est+alfa[3]*x2est+alfa[4]*sqr(x1est)+
        alfa[5]*sqr(x2est)+alfa[6]*x1est*x2est;

x10est:= deltax1*x1est + x1central;
x20est:= deltax2*x2est + x2central;

SQRES:= 0;

for i:= 1 to n do
    begin
        ychfil:= alfa[1]+alfa[2]*x1delinfil+alfa[3]*x2delinfil+
                alfa[4]*sqr(x1delinfil)+
                alfa[5]*sqr(x2delinfil)+alfa[6]*x1delinfil*x2delinfil;
        errofil:= (yfil - ychfil);
        SQRES:= SQRES+sqr(yfil - ychfil);
        SQREG:= SQREG+sqr(ychfil - somay/n);
    end;
(
writeln (lst,' ey          y          ych          x1          x2
erro');
for i:= 1 to n do
    begin
        writeln (lst,eyfil:9:4,' ',yfil:9:4,' ',ychfil:9:4,' ',
                x1fil:6:3,' ',x2fil:6:3,' ',errofil:9:4);
    end;
)

MEDIAX1BS:= 0; VARX1BS:= 0; SOMA1X1BS:= 0; SOMA2X1BS:= 0;
MEDIAX2BS:= 0; VARX2BS:= 0; SOMA1X2BS:= 0; SOMA2X2BS:= 0;
SOMAX1X2BS:= 0; COVX1X2BS:= 0; DMAH:= 0; DMAHBS:= 0;

MEDX1BS:= 0; VX1BS:= 0; S1X1BS:= 0; S2X1BS:= 0;
MEDX2BS:= 0; VX2BS:= 0; S1X2BS:= 0; S2X2BS:= 0;
SX1X2BS:= 0; CVX1X2BS:= 0; MAH:= 0; MAHBS:= 0;

gd:= detect;

```

```

initgraph (gd, gm, ' ');
setcolor (0);

< COORDENADAS DO PONTO ESTACIONARIO VERDADEIRO >
setcolor (4);
line (trunc(x1s-80),trunc(2*x2central-x2s),
      trunc(x1s+80),trunc(2*x2central-x2s));
line (trunc(x1s),trunc(2*x2central-x2s-80),
      trunc(x1s),trunc(2*x2central-x2s+80));

for ibs:= 1 to nbs do
  begin
    SQRESbs:= 0;
    cst:= 0;
    cst1:= 0;
    cst2:= 0;
    X10S:= 0;
    X20S:= 0;
    X10BS:= 0;
    X20BS:= 0;
    AMOSTRA_ERROS;

    ESTIMA_PARAMETROS;

    for i:= 1 to n do
      begin
        ychbsfil:= alfa[1]+alfa[2]*x1delinfil+alfa[3]*x2delinfil+
                   alfa[4]*sqr(x1delinfil)+
alfas[5]*sqr(x2delinfil)+alfas[6]*x1delinfil*x2delinfil;
        errobsfil:= (yfil - ychbsfil);
        end;

        cst1:= 4*alfa[4]*alfas[5];
        cst2:= alfas[6]*alfas[6];
        cst:= cst1-cst2;
        x10s:= ( alfa[3]*alfas[6] - 2*alfa[2]*alfas[5] ) / cst;
        x20s:= ( alfa[2]*alfas[6] - 2*alfa[3]*alfas[4] ) / cst;

<
        for i:= 1 to 6 do
          begin
            writeln (lst,' alfa',i,2,']= ',alfafil:10:4);
          end;
        writeln (lst) ;
        writeln (lst,x10s:10:6,' ',x20s:10:6);
      }

      x10bs:= deltax1*x10s+x1central;
      x20bs:= deltax2*x20s+x2central;

< COORDENADAS DO PT. ESTAC. BS >
setcolor (15);
circle (trunc(x10bs),trunc(x20bs),1);

```

```

WRITELN (REFUGO4,X10S,' ',X20S);

SX1X2BS:= SX1X2BS+X10S*X20S;
S1X1BS:= S1X1BS+X10S;
S2X1BS:= S2X1BS+SQR(X10S);
S1X2BS:= S1X2BS+X20S;
S2X2BS:= S2X2BS+SQR(X20S);

Ysbs:= a1fa[1]+a1fa[2]*x10s+a1fa[3]*x20s+a1fa[4]*sqr(x10s)+
      a1fa[5]*sqr(x20s)+a1fa[6]*x10s*x20s;
( writeln (' nbs= ',nbs:4,' Ysbs= ',ysbs:8:4);
)
end;

CLOSE (REFUGO4);

MEDX1BS:= S1X1BS/NBS;
VX1BS:= (S2X1BS-NBS*SQR(MEDX1BS))/(NBS-1);
MEDX2BS:= S1X2BS/NBS;
VX2BS:= (S2X2BS-NBS*SQR(MEDX2BS))/(NBS-1);
CVX1X2BS:= (SX1X2BS-NBS*MEDX1BS*MEDX2BS)/(NBS-1);

MAH:= VX2BS*SQR((X1S-X1CENTRAL)/DELTAX1-MEDX1BS)-
      2*CVX1X2BS*((X1S-X1CENTRAL)/DELTAX1-MEDX1BS)*
      ((X2S-X2CENTRAL)/DELTAX2-MEDX2BS)+
      VX1BS*SQR((X2S-X2CENTRAL)/DELTAX2-MEDX2BS);

ASSIGN (REFUGO4,'C:\TURBO55\SILVIAMF\X1IX21.DAT');
RESET (REFUGO4);

CONTBS:= 0; CONTBSCOD:= 0;
FOR I:= 1 TO NBS DO
  BEGIN
    READLN (REFUGO4,X10S,X20S);
    MAHBS:= VX2BS*SQR(X10S-MEDX1BS)-2*CVX1X2BS*(X10S-MEDX1BS)*
           (X20S-MEDX2BS)+VX1BS*SQR(X20S-MEDX2BS);
    IF (MAHBS>MAH) THEN CONTBSCOD:= CONTBSCOD+1;

  END;
PERCENTUALBS:= CONTBS/NBS;
repeat until keypressed;
closegraph;
writeln;
writeln;
writeln (' ANGULO= ',ANG:4,' l1= ',l1:6:3,' l2= ',l2:6:3);
writeln (' SIGMA= ',SIGMA:4:2,' Ysest= ',Ysest:8:4);
WRITELN (' X1CENTAL= ',X1CENTRAL:3,' DELTAX1= ',DELTAX1:2);
WRITELN (' X2CENTAL= ',X2CENTRAL:3,' DELTAX2= ',DELTAX2:2);
writeln (' X10est= ',X10est:8:4,' X20est= ',X20est:8:4);
writeln;
writeln (' MEDX1BS= ',MEDX1BS:8:4,' VX1BS= ',VX1BS:8:4);
writeln (' MEDX2BS= ',MEDX2BS:8:4,' VX2BS= ',VX2BS:8:4);
writeln (' CVX1X2BS= ',CVX1X2BS:8:4);
writeln (' CORRELACAO= ',(CVX1X2BS/SQRT(VX1BS*VX2BS)):8:5);

```

```

writeln (' contbsCOD= ',contbsCOD:4,' (%)BS=
',(CONTBSCOD/NBS*100):6:2);
Writeln;
{
writeln (LST);
writeln (LST);
writeln (LST,' ANGULO= ',ANG:4,' I1= ',I1:6:3,' I2= ',I2:6:3);
Writeln (LST,' X1CENTAL= ',X1CENTRAL:3,' DELTAX1= ',DELTAX1:2);
Writeln (LST,' X2CENTAL= ',X2CENTRAL:3,' DELTAX2= ',DELTAX2:2);
writeln (LST,' SIGMA= ',SIGMA:4:2,' Ysest= ',Ysest:8:4);
writeln (LST,' X10est= ',X10est:8:4,' X20est= ',X20est:8:4); }
{ writeln (LST,' MEDX1BS= ',MEDX1BS:8:4,' VX1BS= ',VX1BS:8:4);
writeln (LST,' MEDX2BS= ',MEDX2BS:8:4,' VX2BS= ',VX2BS:8:4);
writeln (LST,' CVX1X2BS= ',CVX1X2BS:8:4);
writeln (LST,' CORRELACAO= ',(CVX1X2BS/SQRT(VX1BS*VX2BS)):8:5);
writeln (LST,' NBS= ',NBS:4,' (%)BS=
',(CONTBSCOD/NBS*100):6:2);
Writeln (LST,' MAH= ',MAH:12:6);
}
read (i);
clrscr;
close (dados);
close (matriz);
CLOSE (REFUGO4);
end.

```

- REFERÊNCIAS BIBLIOGRÁFICAS:

- [1] - BOX, G. E. P. e WILSON, K. B. (1951). On the Experimental Attainment of Optimum Conditions, Journal of Royal Statistical Society, B. v.13, no.1, p. 1-45.
- [2] - BOX, G. E. P. (1954). The Exploration and Exploitation of Response Surfaces: some general considerations and examples. Biometrics. v. , março, p. 16-60.
- [3] - BOX, G. E. P. e HUNTER, J. S. (1957). Multifactor Experimental Designs for Exploring Response Surfaces. Annals of Mathematical Statistical. v.28, p. 195-241.
- [4] - BOX, G.E.P., e DRAPER, N.R. (1959). A basis for selection of a Response Surface Design. JASA. 54, p. 622-654.
- [5] - BOX, G. E. P. ; HUNTER, W. G. e HUNTER, J.S. (1978). Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building. New York, John Wiley & Sons, Inc.
- [6] - EFRON, B. (1979). Bootstrap methods: another look at the Jackknife. Annals of Statis. v.7, no.1, p. 1-26.

- [7] - MARDIA, K. V.; KENT, J.T. e BIBBY, J. M. (1979). Multivariate Analysis. London, Academic Press Inc.
- [8] - FREEDMAN, D. A. (1981). Bootstrapping Regression Models. Annals of Statis. v.9, no.6, nov., p. 1218-1228.
- [9] - BERAN, R. (1982). Estimated Sampling Distributions: The Bootstrap and Competitors. The Annals of Statistics. v.10, no.1, p. 212-225.
- [10] - EFRON, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. SIAM. Philadelphia.
- [11] - DIACONIS, P. D. e EFRON, B. (1983). Computer-Intensive Methods in Statistics. Scientific American. Maio, p.96-108.
- [12] - FREEDMAN, D. A. e PETERS, S. G. (1984). Bootstrapping a Regression Equation: Some Empirical Results. Journal of the American Statistical Association. v.79, no.385, março, p. 97-106.
- [13] - MONTGOMERY, D.G. (1984). Design and Analysis of Experiments. 3ed. New York, John Wiley & Sons, Inc.
- [14] - BOX, G. E. P. e DRAPER, N. R. (1987). Empirical Model-Building and Response Surfaces. New York, John Wiley & Sons, Inc.

- [15] - KHURI, A. I. e CORNEL, J. A. (1987). Response Surfaces: Designs and Analysis. Milwaukee, ASQC.
- [16] - DRAPER, N. R. e SMITH, H. (1981). Applied Regression Analysis. 2ed. New York, John Wiley.
- [17] - CHAVES, A. N. (1985). Bootstrap e Regressão, uma opção aos mínimos quadrados generalizado. Dissertação (Mestrado em Estatística) - IMECC, UNICAMP.