2.2. ANÁLISE ESTATÍSTICA

2.2.1. DADOS OPERACIONAIS

Uma unidade industrial monitorada a partir da análise sistemática individual das variáveis tidas como principais tende a ser de difícil controle, facilmente sujeita a instabilidades decorrentes de variações operacionais, comuns a todas unidades, por diferentes causas.

Contudo, este tipo de monitoramento pressupõe que a qualidade destes dados seja boa, permitindo que se faça uma análise mais abrangente, utilizando ferramentas estatísticas, procurando relacionar as variáveis, tentando estabelecer outras possibilidades que facilitem o controle da unidade, estabilizando-a, melhorando consequentemente a operação, tornando possível sua otimização.

Além da boa qualidade dos dados disponíveis, um passo inicial fundamental para o sucesso deste tipo de análise é saber como os organizar, sendo para tal necessário ter um bom conhecimento da unidade em estudo, evitando misturar dados gerados com características processuais ou produtivas diferentes, o que torna a interpretação dos resultados difícil, além de invalidar qualquer tipo de predição. Uma análise global dos dados disponíveis, constituindo um universo expressivo, mas misturando características operacionais, conforme mencionado, é usualmente menos conclusiva que a análise feita sobre um universo constituído de um número expressivamente menor de dados, mas em que certas características operacionais foram selecionadas.

Evidentemente, para que tal comparação possa ser feita, há a necessidade de que a análise sobre os universos em comparação seja feita, permitindo posteriormente avaliar qual universo é mais representativo do modo que se pretende controlar a unidade.

Apesar da análise pretendida neste estudo utilizar ferramentas estatísticas, o conhecimento do processo é fundamental para que os resultados sejam coerentes, preditivos e conclusivos.

Concluindo, seja qual for o número de universos disponíveis, em que os dados operacionais estejam distribuídos, a análise estatística de todos se faz necessária, para que, com seus resultados, interpretados conjuntamente com as características processuais, operacionais ou produtivas, que diferenciam os universos estudados, seja possível estabelecer de forma segura o modo através do qual se pretende realizar as predições.

2.2.2. MATRIZ DE CORRELAÇÕES

Designada por *Matrix Plots* no software estatístico Minitab, utilizado no desenvolvimento deste estudo, consiste de uma matriz bi-dimensional composta por gráficos correlacionando as diversas variáveis entre si, sendo útil para visualizar de forma imediata as potenciais correlações entre duas variáveis, entre todas as variáveis envolvidas, já que é possível identificar as correlações significativas em um gráfico, economizando tempo (Minitab, 2000).

A matriz de correlações é utilizada para iniciar a análise estatística dos dados históricos dos diversos universos estudados, identificando visualmente as variáveis envolvidas no estudo que se relacionam entre si, conforme mostrado a seguir:

MatrixPlot 'Z1' 'Z2' 'Z3' 'Z4' 'Z5' 'Z6' 'Z7' 'Y2';

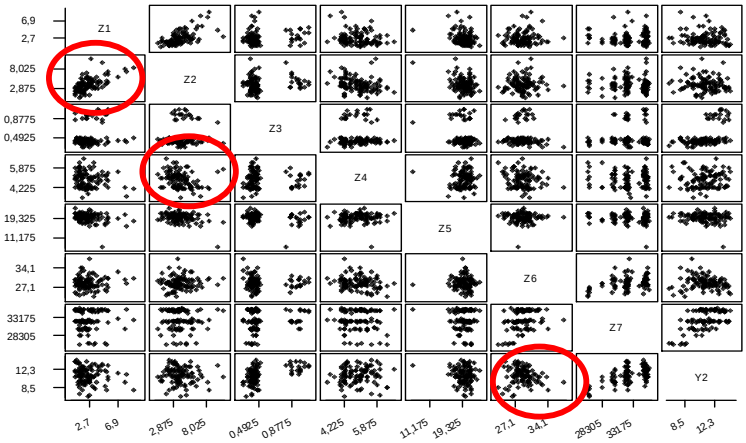

Figura 2.27 – Exemplo de matriz de correlações de análise estatística, indicando as variáveis que se relacionam.

2.2.3. CARTA DE CONTROLE

As matrizes bi-dimensionais de correlação entre as variáveis evidenciam pontos que eventualmente estejam fora de controle, não possibilitando muitas vezes que se detecte visualmente as potenciais correlações.

A carta de controle ou carta de valores individuais de cada uma das variáveis envolvidas permite visualizar pontos que estejam fora de controle, permitindo, por exemplo, melhorar a qualidade da análise via as matrizes de correlação, caso os conjuntos contendo estes pontos sejam eliminados.

Para tanto, é necessário que, após a realização inicial da análise via matrizes de correlações, para todas as matrizes bi-dimensionais em que se julgue haver pontos fora de controle, que a carta de controle da variável com problema seja aberta. Com a carta é possível detectar se a excessiva variação percebida é normal ou não, de acordo com o determinado pelos métodos descritos na literatura, a partir dos dados observados (Montgomery, 2001). Se for considerada normal, ou explicável por algo de conhecimento comum, o conjunto é preservado. Caso contrário, há necessidade de eliminar o conjunto de dados que contenha aquele ponto.

Usualmente percebe-se que a variável é muito constante e que o ponto visualizado como fora de controle está dentro de sua normalidade. É possível também detectar erros de leitura, ou mesmo de digitação; neste caso, se uma correção for possível, o conjunto de dados a que pertence o ponto é mantido.

Por outro lado, ao abrir as diversas cartas de controle das variáveis em que se detectou uma variação excessiva, é possível explicar uma época de operação anormal, facilitando a decisão de eliminar os conjuntos que contenham estes valores anômalos.

Após concluir a análise através das cartas de controle, conforme o descrito, eliminando os conjuntos contendo os referidos pontos, se faz necessário fazer nova análise através da matriz de correlação, que, com segurança, proporcionará uma melhor visualização das eventuais correlações entre as diversas variáveis, permitindo a evolução da análise estatística.

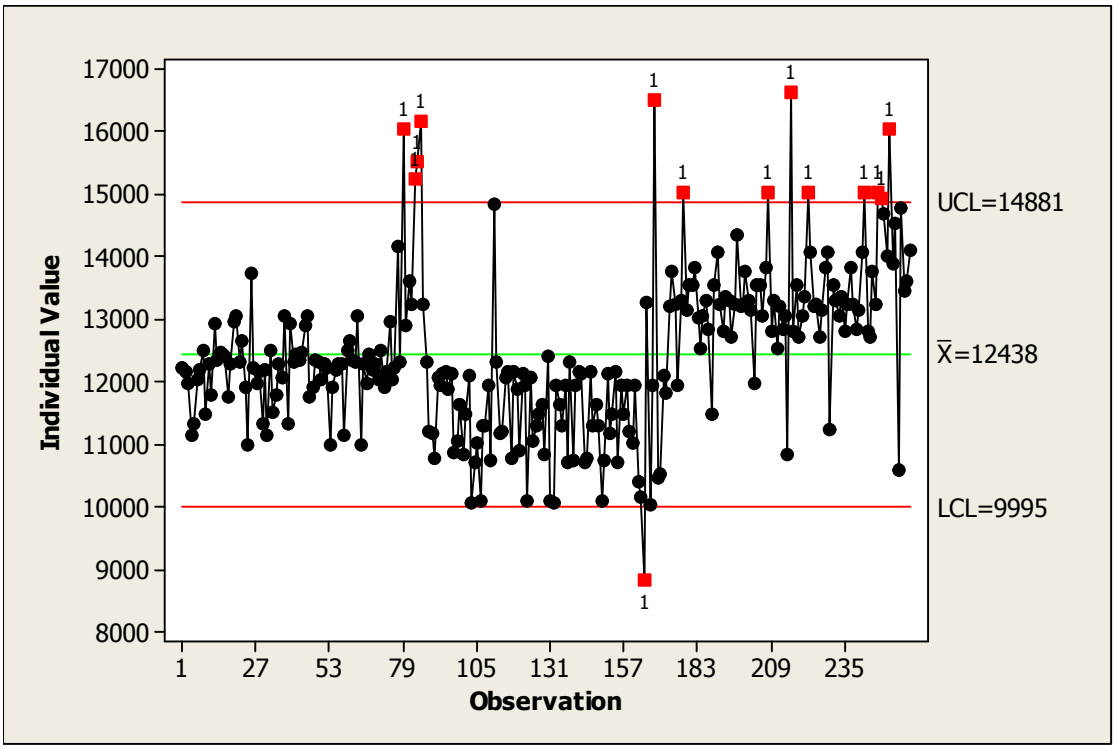

Figura 2.28 – Exemplo de carta de controle de variável envolvida em análise estatística, sugerindo a eliminação de conjuntos em que se evidencia a perda controle.

2.2.4. TESTE DE CORRELAÇÃO DE PEARSON

Compreende a determinação do grau de relação entre duas variáveis, dado pelo coeficiente de Pearson, também chamado de coeficiente de correlação, ou ainda, simplesmente correlação para os pares de variáveis. Este coeficiente de correlação expressa o grau de dependência linear entre duas variáveis. O coeficiente de correlação tem valores entre –1 e +1, sendo negativa quando uma variável diminui com o aumento da outra variável, e positiva quando uma variável aumenta com o aumento da outra.

Para duas variáveis quaisquer sendo testadas, *x* e *y*, calcula-se o coeficiente de correlação de Pearson_ *r*, como segue:

$$r = \frac{\sum (x - x^{o}) \cdot (y - y^{o})}{(n-1) \cdot s_x \cdot s_y} \qquad (2.11)$$

onde,

x^o - média do conjunto de dados da primeira variável;
s_x – desvio padrão do conjunto de dados da primeira variável;
y^o - média do conjunto de dados da segunda variável;
s_y – desvio padrão do conjunto de dados da segunda variável.

Supondo a distribuição normal dos dados, a significância do coeficiente de correlação é testada, via determinação do nível de significância, expresso pelo de valor *p*, para testar as hipóteses nula e alternativa, isto é:

H0: r = 0, não há correlação.
H1: r ≠ 0, há correlação.

A rejeição da hipótese nula será julgada com o seguinte critério:

Correlação fraca 0,05 < p ≤ 0,1
" forte 0,01 < p ≤ 0,05
" fortíssima p < 0,01

O Minitab imprime todos os testes entre os pares de variáveis como uma matriz, indicando numa primeira linha o coeficiente de Pearson e, na segunda linha o valor *p*. Desta forma, pode-se visualizar facilmente quais as variáveis que se relacionam entre si, bem como, comparar as relações entre os diferentes pares de variáveis, como se pode observar no exemplo a seguir:

Correlations: Z1; Z2; Z3; Z4; Z5; Z6; Z7; Y2

	Z1	Z2	Z3	Z4	Z5	Z6	Z7
Z2	0,600						
	0,000						
Z3	0,134	0,153					
	0,179	0,125					
Z4	-0,184	-0,256	0,013				
	0,064	0,009	0,896				
Z5	-0,283	-0,360	-0,187	0,031			
	0,004	0,000	0,060	0,756			
Z6	0,068	0,065	-0,062	-0,200	-0,032		
	0,499	0,515	0,539	0,043	0,749		
Z7	0,005	-0,008	0,121	-0,045	-0,091	0,357	
	0,962	0,935	0,224	0,652	0,364	0,000	
Y2	-0,135	-0,083	0,490	0,067	-0,143	-0,142	0,595
	0,177	0,408	0,000	0,502	0,153	0,153	0,000

Cell Contents: Pearson correlation
P-Value

Figura 2.29 – Exemplo de matriz de teste de correlação de Pearson de análise estatística, indicando as variáveis que se relacionam.

2.2.5. CODIFICAÇÃO DAS VARIÁVEIS

No planejamento de um programa experimental, o executor vê-se usualmente com os seguintes problemas:

- escolher as variáveis de entrada a serem utilizadas no experimento;

- selecionar a faixa de variação, bem como o número de níveis de cada variável, de forma a avaliar adequadamente os efeitos de cada variável sobre a resposta.

Somente ao resolver estes problemas o programa experimental estará definido. Uma questão que dificulta a interpretação da resposta é a diferença de natureza das variáveis de entrada, bem como, no caso das variáveis serem da mesma natureza, diferenças de unidades, ou ainda, diferentes faixas de variação. Dependendo das diferenças descritas, a resposta é de dificílima interpretação ou visualização. Por exemplo, no caso do experimento de uma reação química com três variáveis independentes, temperatura, tempo e pressão, variando entre 120 e 230ºC, 5 a 10 segundos e 5,3 a 8,8 kgf/cm² respectivamente, os níveis mínimos e máximos estariam dispostos nos vértices de um cubo, cuja construção seria difícil, bem como difícil seria sua visualização.

A codificação das variáveis independentes, normalizando os valores das variáveis entre –1 e +1, padronizaria o formato do cubo, facilitando a visualização das respostas.

A utilização de variáveis codificadas no lugar das variáveis de entrada, em sua forma original, como apresentada, facilita a montagem de projetos de experimentos. A codificação remove as unidades de medida dos variáveis, bem como normaliza as dimensões do cubo que expressa suas possíveis variações.

Se expressa, portanto, a codificação das variáveis como:

$$x_i = \frac{2X_i - (X_{iL} + X_{iH})}{(X_{iL} - X_{iH})} \qquad (2.12)$$

onde,

X_i - variável de entrada;

X_{iH} - valor máximo da variável;

X_{iL} - valor mínimo da variável;

x_i - variável codificada (Khuri& Cornell, 1987).

No caso do exemplo relativo ao experimento da reação química com três variáveis, temperatura, tempo e pressão, define-se a temperatura como X_1, sendo $120<X_1<230$ºC, tempo como X_2, sendo $5<X_2<10$ segundos e pressão como X_3, sendo $5,3<X_3<8,8$ kgf/cm². As variáveis codificadas para o exemplo são definidas, portanto como:

temperatura ➔ $x_1 = \frac{2X_1 - 350}{110}$ (2.13)

tempo ➔ $x_2 = \frac{2X_2 - 15}{5}$ (2.14)

pressão ➔ $x_3 = \frac{2X_3 - 14,1}{3,5}$ (2.15)

Observa-se que, para cada uma das variáveis independentes codificadas, quando a variável original de entrada tem o valor mínimo, o valor obtido para a variável codificada é – 1, +1, quando a variável original de entrada tem o valor máximo, e 0, quando a variável original de entrada tem o valor médio.

Existem diversas vantagens em utilizar as variáveis codificadas ao ajustar modelos polinomiais, sendo as principais:

- facilidade computacional e aumento da precisão na determinação dos coeficientes do modelo;

- melhora a interpretação e visualização da estimativa dos coeficientes no modelo;

- possibilita a execução de análises, preservando a confidencialidade de dados operacionais.

2.2.6. REGRESSÃO

2.2.6.1. INTRODUÇÃO

A análise por regressão é uma ferramenta estatística que utiliza as relações entre duas ou mais variáveis, de tal forma que uma variável possa ser predita a partir da outra ou das outras (Netter et alii, 1983).

A regressão pode expressar relações funcionais ou estatísticas. A funcional é expressa por fórmulas matemáticas, em que X_i, sendo i = 1, 2, 3, ..., são as variáveis independentes e Y_j, sendo j = 1, 2, 3, ..., são as variáveis dependentes, determinadas a partir das primeiras.

A relação estatística não é perfeita, pois, em geral, os pontos não se localizam sobre a curva, sendo que, cada ponto, é resultado de uma observação ou ensaio, realizado de forma aleatória. Apesar das relações estatísticas serem muito úteis, não são tão exatas quanto as relações funcionais.

Um modelo de regressão é um meio de expressar duas características essenciais de uma relação estatística:

- a tendência das variáveis dependentes Y_j se relacionarem com as variáveis independentes X_i de um modo sistemático, ou funcional, ou seja, através de fórmulas matemáticas, ou, modelos;

- o conjunto de pontos espalhados ao redor de uma curva de relação estatística.

Como a realidade precisa ser reduzida a proporções manipuláveis, sempre que se estabelece um modelo de regressão, apenas um certo número de variáveis independentes deve ser incluído no modelo de regressão, o que é um ponto importante a ser discutido. Juntamente com este ponto, deve-se configurar a forma funcional do modelo que se pretende utilizar, apesar de que, usualmente, só se pode determinar com precisão este ponto após coleta de dados e análises.

A determinação do modelo de regressão deve garantir a três propósitos básicos na análise de regressão de um determinado fenômeno ou evento em estudo:

- descrição;
- controle;
- predição.

O modelo básico de regressão é o estabelecido somente para uma variável independente e a função de regressão é linear, expressa por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad (2.16)$$

onde,

Y_i – valor da variável de resposta no ensaio ou medida *i*;

$ß_0$ e $ß_1$ – parâmetros de regressão;

X_i – valor da variável independente no ensaio ou medida *i*;

ε_i – termo relativo ao erro aleatório com média $E(\varepsilon_i)=0$ e variância $\sigma^2(\varepsilon_i)=\sigma^2$;

i = 1, 2, 3, ..., n

O modelo descrito é dito simples, linear em parâmetros e linear em variáveis independentes, também denominado de modelo de primeira ordem e a função de regressão é uma linha reta. Os parâmetros de regressão $ß_0$ e $ß_1$ normalmente não são conhecidos, sendo determinados através de dados experimentais ou não experimentais, utilizando métodos como o dos mínimos quadrados, da máxima semelhança, etc..

A análise de regressão múltipla é uma das mais largamente utilizadas de todas ferramentas estatísticas. Quando duas variáveis independentes X_1 e X_2 são utilizadas, o modelo é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad (2.17)$$

onde,

Y_i – valor da variável de resposta no ensaio ou medida *i*;

$ß_0$,$ß_1$ e $ß_2$ – parâmetros de regressão;

X_{i1} e X_{i2} – valores das variáveis independentes no ensaio ou medida *i;*

ε_i – termo relativo ao erro aleatório com média $E(\varepsilon_i)=0$ e variância $\sigma^2(\varepsilon_i)=\sigma^2$;

i = 1, 2, 3, ..., n

A função de regressão deste modelo é um plano e, analogamente ao modelo de primeira ordem, os parâmetros de regressão $ß_0$,$ß_1$ e $ß_2$,

igualmente desconhecidos, são determinados pelos mesmos métodos descritos anteriormente.

Para a análise de regressão utilizando um modelo de múltiplas variáveis, define-se o modelo geral de regressão, dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots + \beta_{p-1} X_{ip-1} + \varepsilon_i \qquad (2.18)$$

onde,

Y_i – valor da variável de resposta no ensaio ou medida i;

$ß_0$,$ß_1$, $ß_{2\ ,\ \ldots\ldots\ldots}$, $ß_{p-1}$ – parâmetros de regressão;

$X_{i1,}$ $X_{i2,}$ ....., X_{ip-1} – valores das variáveis independentes no ensaio ou medida i;

ε_i – termo relativo ao erro aleatório com média $E(\varepsilon_i)=0$ e variância $\sigma^2(\varepsilon_i)=\sigma^2$;

i = 1, 2, 3, ..., n

A função de regressão deste modelo é um hiperplano, que é um plano em mais de duas dimensões, representada graficamente, para determinadas variáveis fixas, através da superfície de resposta. Analogamente ao modelo de primeira ordem, os parâmetros de regressão $ß_0$, $ß_1$, $ß_2$, ..., $ß_{p-1}$, igualmente desconhecidos, são igualmente determinados pelos mesmos métodos descritos anteriormente (Netter et alii, 1983).

2.2.6.2. DETERMINAÇÃO DO MELHOR SUBCONJUNTO DE REGRESSÃO

Como o objetivo de uma análise por regressão é o de estabelecer um modelo que descreva, controle e faça predições sobre o fenômeno em estudo,

um problema que se impõe é o de estabelecer o conjunto adequado de variáveis independentes a serem consideradas no modelo.

Para um fenômeno qualquer em estudo, o passo inicial é o de observar as variáveis independentes que potencialmente o influenciam, o que pode usualmente resultar num número elevado; inicia-se então um processo de seleção, eliminando as variáveis independentes excedentes de acordo com o seguinte critério:

- não são realmente fundamentais para o fenômeno em estudo;

- estão sujeitas a grandes erros de medida;

- podem efetivamente duplicar outra variável independente.

Tipicamente, o número de variáveis independentes remanescentes após uma seleção preliminar, como anteriormente descrita, continua a ser elevado. Além do mais, muitas destas variáveis são potencialmente correlacionáveis, piorando a qualidade do modelo, quanto à sua capacidade de executar predições. Trabalhar com modelos com elevado número de variáveis independentes eleva desnecessariamente o custo, pois envolve número excessivo de medições e análises, aumentando ou potencializando a probabilidade de ocorrência de erros, que interferem na exatidão do modelo.

Apesar de haver a necessidade de eliminar variáveis, deve-se tomar cuidado, contudo, para que variáveis explicativas, fundamentais para a exatidão do modelo, não sejam desnecessariamente eliminadas, o que pode danificar seriamente a capacidade esclarecedora do modelo, conduzindo ainda a determinação de coeficientes de correlação afetados por interferências, que resultam em respostas fracas, além de predições falhas.

A questão é então saber como diminuir o número de variáveis independentes, mantendo as qualidades do modelo, executando, portanto a boa seleção de variáveis independentes. Este subconjunto de variáveis independentes precisa ser pequeno o suficiente para redução de custos e facilidade de análise, e, ao mesmo tempo, grande o suficiente para garantir as qualidades do modelo quanto à descrição, controle e predição.

Como as razões para execução da análise por regressão variam, não existe um subconjunto que seja o melhor para todos usos. Para um determinado fenômeno em estudo, é comum achar vários subconjuntos que apresentem igual desempenho, mas a escolha do subconjunto a ser utilizado num modelo de regressão precisa ser feita com base em considerações adicionais. O processo de seleção é, e deveria ser, pragmático, apesar de ser importante saber não desprezar fundamentais julgamentos subjetivos. Deve-se sempre evitar julgamentos por razões mecânicas, como, por exemplo, eliminar uma variável independente porque, na amostra analisada, apresentou uma estreita faixa de variação, tornando-se estatisticamente sem significado, o que não é verdade.

O procedimento descrito na referência (Netter et alii, 1983), identificado como *seleção de todas regressões possíveis*, requer um exame de todos os possíveis modelos, envolvendo as potenciais variáveis independentes _*X*, identificando os subconjuntos “bons”, de acordo com um critério estabelecido, a ser descrito a seguir. Inicia-se a análise pelo modelo de regressão sem variáveis independentes *X*, seguindo-se dos modelos de regressão com uma variável independente X_i ($X_1, X_2, X_3, X_4, ...$), e, depois destes, os modelos de regressão com duas variáveis independentes X_i (X1 e X2, X1 e X3, X1 e X4, ..., X2 e X3, X2 e X4, X3 e X4, ...), e assim por diante.

Diferentes critérios de comparação podem ser utilizados no procedimento *seleção de todas regressões possíveis*, mas conforme o usualmente indicado

(Hocking, 1976; Netter et alii, 1983, Breyfogle, 1999), bem como aplicado no software utilizado neste estudo (Minitab, 2000), três critérios são mais comuns e serão descritos com maiores detalhes.

Critério R^2_P

Define-se o *coeficiente de determinação* R^2_p como sendo a porcentagem de pontos ajustados ao modelo de regressão, com relação ao universo total de dados. É utilizado para selecionar um ou vários subconjuntos de variáveis independentes, onde o número de parâmetros no modelo analisado é indicado pelo índice *p*, significando, portanto que o modelo contém *p* componentes, contando com a constante independente da variável independente, ou seja, *p-1* variáveis independentes, sendo expresso por:

$$Rp^2 = \frac{SSRp}{SSTO} = 1 - \frac{SSEp}{SSTO} \qquad (2.19)$$

onde,

SSE_P - mede a variação da variável dependente Y, quando o modelo de regressão com p –1 variáveis independentes é utilizado;

$SSTO$ - mede a variação (ou a imprecisão) da variável dependente Y, quando nenhum modelo de regressão é utilizado;

SSR_P - é a diferença entre SSE_P e $SSTO$, também denominado resíduo (Netter et alii, 1983).

Como R^2_P é a razão das somatórias dos quadrados e o denominador é constante, porque independe do modelo de regressão, e como SSE_P diminui com o aumento de variáveis independentes, R^2_P aumenta com o acréscimo de variáveis (Netter et alii, 1983). O maior valor possível de R^2_P é 1, o que ocorre

somente quando não houver resíduo, ou seja, quando o modelo de regressão proposto estiver totalmente ajustado ao universo total de dados (Farias, 2004).

O critério de determinar o *coeficiente de determinação* R^2_p, que é o mais comumente utilizado na análise de seleção de melhor subconjunto, consiste não em maximizar R^2_P, mas, principalmente saber qual é o ponto em que a inclusão de novas variáveis no modelo de regressão traz acréscimos desprezíveis em R^2_P. Freqüentemente, o ponto ótimo é atingido quando um número limitado de variáveis independentes é considerado no modelo de regressão.

Construindo um gráfico de R^2_P versus *p*, a região próxima ao ponto ótimo mostra basicamente que R^2_P não varia com *p*, nada significando aumentar *p*, significando sim, ser possível reduzir o número de variáveis independentes sem, contudo, alterar R^2_P, que continua próximo ao valor máximo. Continuando a reduzir o número de variáveis independentes, confirma-se esta variação desprezível de R^2_P, até que, para um determinado valor de *p*, abaixo do qual se verifica uma drástica redução de R^2_P; o número de variáveis independentes *p*, em que se verifica este "joelho" é o verdadeiro ponto ótimo, sendo frequentemente utilizado para definir o modelo de regressão (Hocking, 1976; Netter et alii, 1983).

Critério R^2_a

Define-se o *coeficiente de determinação ajustado* R^2_a, como sendo a porcentagem de pontos ajustados ao modelo de regressão, com relação ao número total de pontos da amostra.

Como no critério R^2_P o valor máximo nunca diminui com o aumento de *p*, a *análise do coeficiente de determinação ajustado* _ R^2_a é executada, levando

em consideração o número de variáveis independentes, através do grau de liberdade, sendo expressa por:

$$Ra^2 = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSEp}{SSTO} \qquad (2.20)$$

Sendo o erro médio quadrático das variáveis independentes do modelo_ MSEp expresso por:

$$MSEp = \frac{SSEp}{n-p} \qquad (2.21)$$

sendo,

$(n-p)$ – número de graus de liberdade com que o erro é estimado.

Expressa-se o *coeficiente de determinação ajustado* _ R^2_a por:

$$Ra^2 = 1 - \frac{MSEp}{\frac{SSTO}{n-1}} \qquad (2.22)$$

Pode-se observar que R^2_a aumenta somente se MSE_p diminuir, já que $SSTO/(n-1)$ é constante para um dado número de observações de Y na amostra. Portanto, R^2_a e MSE constituem um critério equivalente. O mínimo MSE_p pode verdadeiramente aumentar com o aumento de p, o que resulta na diminuição de SSE_p, que se torna tão pequeno que não é suficiente para compensar a perda devido ao aumento dos graus de liberdade.

O critério consiste, portanto, em determinar o subconjunto de variáveis independentes X, que minimize MSE_p, ou então um ou vários subconjuntos para os quais MSE_p é tão próximo do mínimo que, considerando mais variáveis no modelo de regressão não faça nenhuma diferença (Hocking, 1976; Netter et alii, 1983).

Critério C_p

C_p expressa a relação entre o erro quadrático médio total e as m variáveis ajustados para cada um dos diversos subconjuntos de variáveis independentes constituintes dos modelos de regressão (Hocking, 1976).

Assume-se que o modelo que incluir todas $P-1$ potenciais variáveis independentes _ X, cuidadosamente escolhidas de tal forma que o erro quadrático médio _ MSE das $P-1$ variáveis ajustadas (MSE_m) é uma estimativa de σ^2.

Para estas condições, C_p é expresso por:

$$Cp = \frac{SSEp}{MSEm} - (n - 2p) \qquad (2.23)$$

O critério de utilização do C_p consiste em procurar identificar os subconjuntos de variáveis independentes _ X para os quais:

- o valor de C_p seja mínimo (indicando que o modelo está ajustado, por ter variância mínima);

- o valor de C_p seja próximo a *p* (número de variáveis independentes do modelo de regressão).

Subconjuntos com valores mínimos de C_p têm um valor mínimo de erro quadrático médio total. Quando o valor de C_p é próximo ao valor de *p* a interferência no modelo de regressão é pequena (Hocking, 1976; Netter et alii, 1983).

Pode ocorrer algumas vezes que o modelo de regressão baseado no subconjunto de variáveis independentes _ *X* com o mínimo valor de C_p apresente alguma interferência. Neste caso, deve-se preferir utilizar um modelo de regressão, baseado em um número pouco maior de variáveis independentes _ *X*, para o qual o C_p seja ligeiramente maior, mas que não envolva significativa interferência, como é o caso da análise complementar executada pelo software Minitab (Minitab, 2000), utilizado no desenvolvimento deste estudo, que apresenta a determinação da variância _s, raiz quadrada do erro quadrático médio _ MSE, consistindo, portanto num critério complementar de análise (Hocking, 1976).

Para ilustrar o descrito neste item, apresenta-se abaixo um quadro de saída do Minitab, contendo todos os conceitos descritos.

Best Subsets Regression: Y2 versus Z^2; Z^3; Z^4; Z^5; Z^6; Z^7; Z^8

Response is Y2

					Z^2	Z^3	Z^4	Z^5	Z^6	Z^7	Z^8
Vars	R-Sq	R-Sq(adj)	C-p	S							
1	40,8	40,2	86,7	1,2298						X	
1	22,4	21,6	142,1	1,4079	X						
2	56,3	55,4	41,8	1,0620					X	X	
2	56,2	55,2	42,2	1,0636	X					X	
3	68,9	67,9	5,8	0,90071	X				X	X	
3	57,9	56,5	39,1	1,0485	X	X				X	
4	70,0	68,7	4,5	0,88966	X	X			X	X	
4	69,1	67,7	7,2	0,90289	X				X	X	X
5	70,5	68,8	5,1	0,88751	X	X		X	X	X	
5	70,3	68,6	5,7	0,89032	X	X			X	X	X
6	70,8	68,8	6,1	0,88748	X	X		X	X	X	X
6	70,5	68,5	7,1	0,89237	X	X	X	X	X	X	
7	70,8	68,5	8,0	0,89204	X	X	X	X	X	X	X

Figura 2.30 – Exemplo de saída do Minitab para seleção de subconjunto de variáveis independentes para modelo de regressão.

2.2.6.3. ESTABELECIMENTO DO MODELO DE REGRESSÃO

Definido o melhor subconjunto de variáveis independentes *X*, seja utilizando qualquer um dos critérios descritos no item anterior, ou, conforme o realizado pelo software Minitab, utilizado no desenvolvimento deste estudo, ajusta-se o modelo de regressão múltipla para o mesmo universo de dados, definindo-se as variáveis do modelo, executando a diagnose da análise, o que se faz principalmente pela visualização da probabilidade de erro que cada variável introduz, significando que quanto menor seu valor, ou seja, quanto mais próximo de zero, melhor, significando que a baixa probabilidade de risco de rejeitar a hipótese de que o coeficiente é nulo, quando ele for verdadeiramente nulo, ou ainda, baixos valores de *p*, como será referenciado na análise indica que o coeficiente realmente existe (Soares & Siqueira, 2002).

A indicação do coeficiente de determinação ajustado (R^2_a), conforme o detalhado em 2.2.6.2, é um critério para avaliar o quanto, em termos de porcentagem de pontos ajustados, o modelo regressão ajustado explica das variações observadas, com relação ao número total de pontos da amostra. Há ainda a verificação gráfica da distribuição dos resíduos, diferença entre as respostas calculadas pelo modelo e as medidas.

Por fim, identificam-se os pontos que estão fora do ajuste, em inglês *outliers*, definidos através do termo distância, que, se superior a 2, devem ser eliminados, o que usualmente requer nova análise, visando verificar se o subconjunto selecionado na análise inicial é confirmado.

Repete-se o ajuste do modelo de regressão, mantendo o mesmo subconjunto de variáveis independentes *X,* ou, caso este não tenha sido mantido, o novo subconjunto selecionado, repetindo o procedimento descrito neste item (Minitab, 2000).

Definido o modelo de regressão múltipla, executa-se um teste na instalação em estudo, dentro da mesma faixa de operação dos dados históricos, com o propósito de validar o modelo, definindo-se previamente em que ponto as variáveis independentes devem ser ajustadas, visando melhor confirmação do modelo estabelecido. O modelo estará validado se os valores obtidos para a variável dependente _ *Y*, calculados através do modelo, estiverem no intervalo de confiança estipulado para *Y*.

Da mesma forma do que foi apresentado para o item anterior, também visando ilustrar o descrito neste item, apresenta-se abaixo uma saída do Minitab, contendo o descrito.

Regression Analysis: Y2 versus Z^2; Z^3; Z^5; Z^6; Z^7

The regression equation is
Y2 = 10,5 - 0,577 Z^2 + 1,30 Z^3 - 0,384 Z^5 - 1,89 Z^6 + 2,09 Z^7

Predictor	Coef	SE Coef	T	P
Constant	10,5002	0,2506	41,90	0,000
Z^2	-0,5766	0,2676	-2,15	0,034
Z^3	1,3003	0,1720	7,56	0,000
Z^5	-0,3838	0,3919	-0,98	0,330
Z^6	-1,8948	0,3003	-6,31	0,000
Z^7	2,0920	0,1829	11,44	0,000

S = 0,8836 R-Sq = 72,3% R-Sq(adj) = 70,8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	185,414	37,083	47,49	0,000
Residual Error	91	71,051	0,781		
Total	96	256,465			

Source	DF	Seq SS
Z^2	1	1,986
Z^3	1	73,355
Z^5	1	2,802
Z^6	1	5,140
Z^7	1	102,131

Unusual Observations

Obs	Z^2	Y2	Fit	SE Fit	Residual	St Resid
3	-0,39	9,5000	8,0207	0,3914	1,4793	1,87 X
17	0,55	8,1000	10,0817	0,2429	-1,9817	-2,33R
18	-0,41	9,0000	10,9212	0,1044	-1,9212	-2,19R
59	0,05	11,6000	9,8128	0,2058	1,7872	2,08R
70	-0,29	6,6000	8,4051	0,2464	-1,8051	-2,13R
74	-0,29	11,5000	9,7139	0,1396	1,7861	2,05R
88	0,82	13,3000	13,4680	0,5781	-0,1680	-0,25 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Figura 2.31 – Exemplo de saída Minitab para análise do modelo de regressão múltipla ajustado para as variáveis independentes selecionadas.

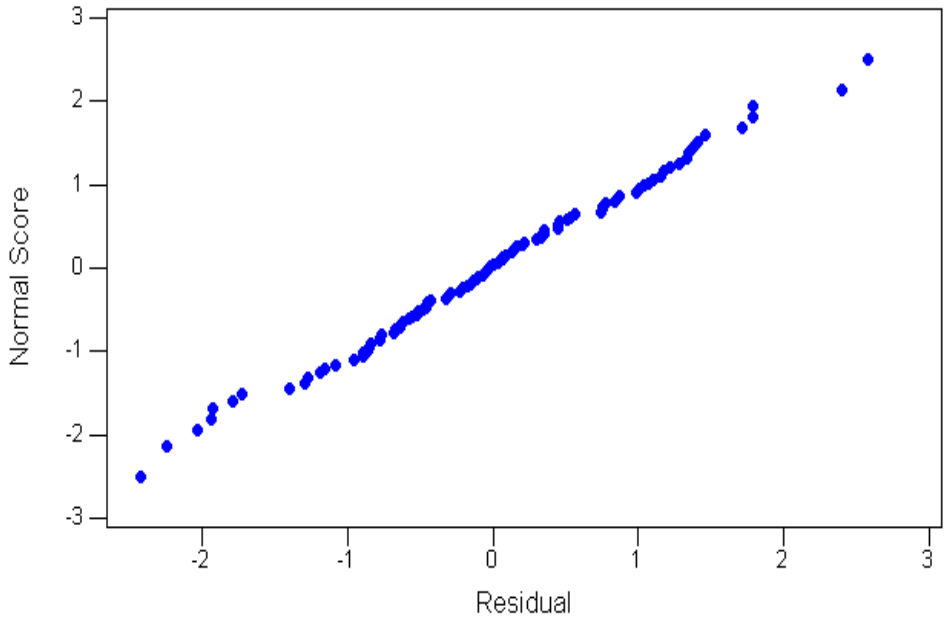

Figura 2.32 Exemplo de saída Minitab para gráfico dos resíduos normalizados do modelo de regressão múltipla ajustado.

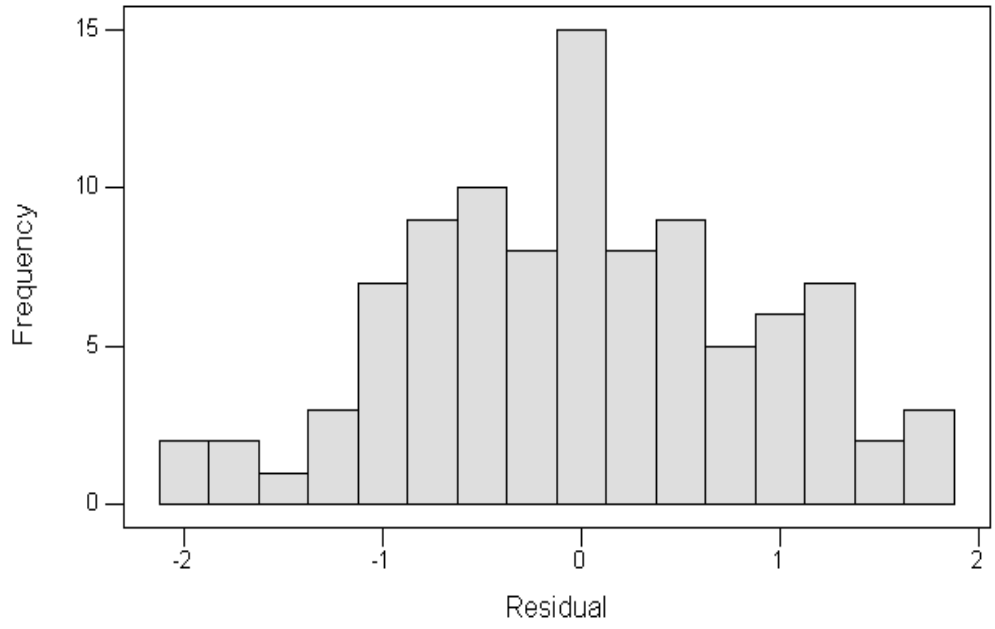

Figura 2.33 – Exemplo de saída Minitab histograma dos resíduos do modelo de regressão múltipla ajustado, para análise estatística realizada no presente estudo.