# Técnicas de Apoio alternativas, a serem utilizadas na ADM – Análise Discriminante Múltipla, no método do Cluster – Agrupamento e no Trato do Banco de Dados Interativo

*Autores*: **Istvan Karoly Kasznar, *PhD***
*Professor Titular da FGV e*
*Presidente da IBCI*

**Bento Mario Lages Gonçalves, *MSc***
*Consultor Sênior da IBCI*

---

São inúmeras as técnicas que se podem utilizar no estudo de grupos característicos, de agrupamentos particulares e de dados que permitem discriminar entre categorias, para balizar convenientemente as decisões que geram técnicas de Classificação e Análise de Potencial.

Em função da sua diversidade e riqueza, procuraremos desenvolver um conjunto básico e fundamental de análises estatísticas e econométricas, relativamente simples, associando-os a dados utilizados e apresentados pelo PEEM – Programa de Estudos dos Estados e Municípios, da Escola Brasileira de Administração Pública e de Empresas, da Fundação Getúlio Vargas, e do banco de dados da IBCI – Institutional Business Consultoria Internacional.

Fundamentalmente, procuraremos dispor de análises quantitativas, apoiadas por análises descritivas, que permitam concluir acerca do processo evolutivo, tendencial, correlacional e de peso ou importância de cada uma das variáveis uma a uma e de todas as variáveis interligadas entre sí, estas no caso das regressões múltiplas.

Com vistas a uma configuração conveniente deste processo de pesquisa, procuraremos analisar a classificação através de técnicas variadas e que serão integradas, para comporem um conjunto coeso e bem correlacionado.

Desta forma, utilizaremos o ADM, o Cluster e o Sistema Interativo que sobrepõe números e relações estatísticas, para então dispor nominalmente de uma classificação por categorias e classes de dados.

Completando o processo, faremos segundo a necessidade projeções lineares, logarítmicas e exponenciais, procurando escolher, entre elas, aquelas que forem as mais convenientes. Neste caso, originalmente, as regressões múltiplas estudam as relações para 20 (vinte) produtos e serviços, ademais das dimensões que contribuem para classificação.

Como há sutilezas na montagem, gestão e disponibilização desta modelagem, o grupo que estudou e fez este *case-study* está empenhou-se em *pari - passu* otimizar e aperfeiçoar esta metodologia .

Estas técnicas possivelmente envolverão análises seqüenciais, assim como também o estudo dos dados , através das suas taxas e tendências.

Trabalhar-se-á com quadros geradores de produtos mensuradores de Classificação e resultados e procurar-se-á utilizar critérios que possuam bases de decisão vinculadas à prática empresarial, representando cortes tecnicamente aceitáveis.

Do ponto de vista estatístico, procurar-se-á estudar os índices de correlação $R_2$, quando aplicáveis.

As técnicas a serem usadas envolverão a utilização de cálculos de desvio padrão, média, variância, regressão linear, regressão múltipla e cálculos de funções projetivas (no caso de produtos e serviços) que melhor acomodem efetivamente a tendência de evolução de variáveis selecionadas, como as estatísticas $F$ e as co-variâncias, entre outros.

## Análise e Estudo de Quadros e Gráficos

Na análise e no estudo das tabelas e dos gráficos que decorrerão, procuraremos essencialmente interpretar os dados globais de ordens macroeconômicas, microeconômicas, municipal e de produtos e serviços. Em função dos resultados obtidos, dada a visualidade maior obtida através de gráficos, procuraremos explicitar os processos dinâmicos que ocorrem caso a caso.

Pretenderemos utilizar tabelas resumidas e auto-interpretativas, e gráficos em duas dimensões e quando aplicável, tridimensionais. Também deveremos utilizar gráficos de barras, com linhas de grade e marcadores de dados tridimensionais, que permitem visualizar, com

facilidade, processos evolutivos especialmente no que diz respeito, neste caso, às dimensões estudadas.

Diversos gráficos de linhas também poderão ser utilizados.

Gráficos de barra poderão também mostrar especificamente índices de participação básicos, em alguma variável específica.

Em função das análises relacionadas a cálculos de desvio-padrão, variâncias e desvios médios, também procuraremos apresentar variados gráficos de dispersão.

Considerando a importância dos picos e dos fundos de vale no processo de interpretação dos ciclos econômicos que ocorrem em cada município, concentraremos possívelmente nossas atenções também em gráficos de máximos e mínimos, nos quais a rota do PIB, também dada pela taxa de crescimento real do PIB, será discutida e analisada cuidadosamente. Os máximos e mínimos, juntados entre diversos pontos de máximos e de mínimos, poderão indicar corredores e tendências específicas de evolução da economia no curto e no médio prazo.

## Instrumentos e Relações Estatísticas

Entre os diversos tipos de análises estatísticas que estudaremos, figuram aquelas que, em geral, são mais utilizadas em análise econômica.

Vamos analisar uma a uma.

Apresentamos a seguir os dados estatísticos, os instrumentos econométricos e as relações.

### 1 - Desvio Padrão

O desvio padrão produz uma estimativa da amplitude em que os valores estão dispersos em relação ao valor médio de uma série de variáveis, logo, em relação à média.

O desvio padrão assume que seus argumentos são uma amostra da população. Se os seus dados representam uma população inteira, o que não será o caso, pois lidar-se-á com 1027 municípios entre os 5.507 existentes no Brasil em 1997, seria necessário calcular um desvio

padrão baseado na amplitude plena com que os valores estão dispersos em relação ao valor médio.

O desvio padrão cuja fórmula será utilizada é:

$$\theta = \frac{\sqrt[m]{\sum X^2 - \left(\sum X\right)^2}}{n.(n-1)}$$

Analisaremos a **média** dos números dados.

A **mediana** corresponde ao número no centro de um conjunto de números, isto é, metade dos números ocorre acima da mediana e a outra metade está abaixo.

**2 - Média**

Procuraremos estudar a média dos argumentos, $\mu$. A média corresponde à somatória dos valores analisados, dividido pelo número de freqüências encontradas:

$$\mu = \frac{\sum_{i=1}^{n} Xi}{n}$$

**4.3.3 - Máximo**

Visa-se estabelecer através de um valor **máximo** o maior número, que consta da lista de variáveis existente. Isto fornece a idéia de um extremo.

**4 - Mínimo**

Visa-se pelo meio de um valor mínimo determinar o menor número da lista de variáveis disponíveis.

Máximos e mínimos são estudados quando fazem-se intervalos de confiança e estudam-se valores extremos, que enquadram todo o restante do conjunto de valores analisados.

**5 - Projeção linear**

Pretendemos caso compatível com nossos propósitos, utilizar o método dos mínimos quadrados para calcular uma linha reta, que melhor se ajusta aos dados disponíveis e produz uma matriz que descreve a evolução de uma linha.

A equação para uma linha é:

$$Y = m_1 X_1 + m_2 X_2 + \ldots + b$$

ou

$$Y = m_i X_i + b$$

quando $i = 1, 2, \ldots n$

Nestas equações, o valor $y$ depende de $X$ e é uma função dos valores $X$ independentes.

Os valores $M$ são coeficientes técnicos correspondentes a cada valor de $X$ e de $B$, em um valor constante.

É importante salientar que $y$, $X$ e $M$ podem ser vetores.

Os valores de erro padrão dos coeficientes são as expressões anteriores que se correlacionarão a $m_1, m_2, \ldots mn$.

**6 - Valor do Erro Padrão**

**EBB** corresponde ao **valor de erro padrão** e da constante **B**.

**7 - Coeficiente de Correlação $R^2$**

$R^2$ **é coeficiente de correlação** também conhecido como de determinação.

O $R^2$ compara valores de $y$ estimados e realizados. Seu valor encontra-se entre 0 e 1. Quando $R^2$ é igual a 1, existe perfeita correlação da amostra. Isto é, há uma resposta direta e imediata entre as variáveis analisadas. Não há diferença entre o valor que se estima de $y$ e o valor efetivamente obtido de $y$.

No outro extremo, sendo o coeficiente de correlação igual a 0, a equação de regressão não é considerada útil para a previsão de um valor de $y$.

O coeficiente de correlação mede a associação normalizada e mostra o valor que oscila entre -1 e +1. Uma correlação positiva mostra que as variáveis se movem em uma mesma direção, e uma correlação negativa mostra que as variáveis se movem em direções opostas.

Por outro lado, qualquer uma das duas variáveis estatisticamente independente terá sempre um coeficiente de correlação igual a 0. A correlação $R^2$ tem normalmente um valor alto na análise dos modelos de regressão, nos quais uma relação causal entre uma variável dependente $y$ e uma variável independente $X$ é muito clara e específica. Portanto, a correlação é interpretada muito mais como uma medida de correlação entre variáveis diversas.

**EPY** corresponde ao erro padrão da estimativa $y$**.**

**8 - Distribuição Estatística F**

O $F$ estatístico ou o **valor observado** $F$ é utilizado para a determinação da relação observada entre as variáveis dependentes e independentes. Procura verificar se esta relação é tal que ela ocorre por acaso ou não.

**9 - Graus de Liberdade**

$GL$ mede os **graus de liberdade**. Utilizam-se os graus de liberdade para compreender os valores críticos de $F$ numa tabela estatística.

Comparam-se os valores encontrados nessa tabela com a estatística $F$, produzidos através da projeção linear para determinação de um grau de confiança para o modelo estudado.

**10 - Soma da Regressão dos Quadrados**

**SQREG** corresponde à soma de regressão dos quadrados.

**11 Soma dos Resíduos dos Quadrados**

**SQRESID** corresponde à soma residual dos quadrados.

Ao desenvolvermos a análise através do método da **regressão** e dos **mínimos quadrados**, é importante que percebamos certos cálculos envolvidos na obtenção da inclinação e da interseção das funções.

A inclinação de **M** será igual a:

$$m = \frac{y_2 - y_1}{X_2 - X_1}$$

A equação da linha reta é:

$$y = mx + b$$

A precisão da linha calculada pela projeção linear depende do grau de dispersão, logo de variabilidade dos seus dados. Quando mais lineares forem os dados, mais preciso será o modelo de projeção linear.

Quanto menos lineares forem os dados e as formações entre eles, mais será necessário sair de projeções lineares e entrar em projeções polinomiais de segundo, terceiro, quarto, quinto ou mais graus; ou ainda trabalharem-se funções logarítmicas ou exponenciais e outras.

A função de projeção linear utiliza o método dos mínimos quadrados para a determinação e melhor ajuste para os dados. Quando existe apenas uma variável independente, os cálculos de **m** e de **b** podem ser baseados e determinados através das fórmulas seguintes:

$$m = \frac{N\sum(XY) - \left(\sum X\right)\left(\sum y\right)}{N\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$b = \frac{\left(\sum y\right)\left(\sum X^2\right) - \left(\sum X\right) - \left(\sum Xy\right)}{N\left(\sum\left(X^2\right)\right) - \left(\sum X\right)^2}$$

As funções de ajuste de curva e de linha dadas por projeções lineares e projeções logarítmicas oferecem, em geral, possibilidades de se calcular uma linha reta ou uma curva exponencial, conveniente e corretas, que se ajustem aos dados estatísticos disponíveis.

Em função disso, é preciso que o analista e os estatísticos venham a decidir-se acerca de qual dos dois resultados melhor se adaptam aos seus dados e às suas necessidades.

Para vermos se a linha ou uma curva calculada ajusta-se aos dados, poderemos também reforçar as análises através de cálculos de tendência e de cálculos de crescimento.

Estas funções podem produzir matrizes de valores **Y** previstos ao longo de uma linha ou de uma curva em seus pontos de dados reais.

Também facilita um processo de comparação de valores previstos com valores reais.

Nas análises de regressão, é importante também que venhamos a calcular, para cada ponto, a diferença dos quadrados entre os valores $y$, estimados para um ponto específico e seu valor $y$ real. A soma dos quadrados das diferenças é denominada de soma residual dos quadrados. Por isso, procuraremos calcular a soma dos quadrados das diferenças entre os valores $y$ reais e a média dos valores de $y$. A esta chamamos de soma total dos quadrados que corresponde, no fundo, à regressão da soma dos quadrados, adicionada da soma residual de quadrados. Quanto maior for a comparação entre a soma residual de quadrados e a soma total de quadrados, menor deverá ser o valor do coeficiente de relação que denominamos de $R^2$ e que representa um indicador do grau de perfeição com o qual a equação resultante da análise de regressão explica as relações entre as variáveis.

## 12 Regressão Linear Múltipla

A **Regressão Linear Múltipla** é especialmente importante, pois permite que se estime o valor de uma variável com base num conjunto de outras variáveis. Quanto mais significativo for o peso de uma ou um conjunto de variáveis dado o universo de disponibilidade de variáveis maior, tanto mais se poderá dizer que alguns fatores afetam mais uma variável especificamente procurada, do que outros.

A equação de regressão múltipla pode ser dada por uma equação do tipo:

$$y = M_0 + M_1X_1 + M_2X_2 + M_3X_3 + M_4X4 + B$$

Onde $B$ corresponde a um coeficiente técnico fixo, a um valor de base a partir do qual começa $y$**.** Os $m$ correspondem a coeficientes técnicos atrelados às variáveis independentes e, neste caso, constata-se que $X_1, X_2$**,** e $X_3$ correlacionam-se positivamente com $y$**,** enquanto $X_4$ relaciona-se de forma negativa, logo inversa, com $y$. A regressão múltipla não utilizada em nossos estudos, embora mereça consideração, já que é uma ferramenta amplamente usada.

**13 - Estatística t.**

O valor estatístico **t** é um outro teste hipotético que tem por objetivo determinar se cada coeficiente de inclinação é significativo na avaliação do valor estimado de uma variável específica.

Isso corresponde a uma expressão do tipo:

$$t = \frac{m^4}{SE_4}$$

**14 - Variância** $(\theta)^2$

Uma **variância** corresponde a uma estimativa de dispersão de uma população, tendo por base uma amostra oferecida como argumento. A variância parte do princípio de que seus

argumentos correspondem a uma amostra da população. A expressão mais importante é dada pela seguinte fórmula:

$$\theta^2 = \frac{N\sum X^2 - \left(\sum X\right)^2}{N(N-1)}$$

Os desvios são conhecidos como distâncias. O desvio padrão é a raiz quadrada da variância. A rigor, se fizermos um modelo de regressão relacionado a uma série de pontos que podem formar uma curva ou uma função linear, o que se busca é uma minimização das distâncias entre os pontos da linha ou da curva. Então, busca-se a minimização dos desvios. O ***desvio padrão*** é o desvio médio que ocorre com maior freqüência, regularidade e em maior número para uma dada amostra específica de pontos ou variáveis.

### 15 - Covariância $(Cov)$

Uma **co-variância** é uma média da associação linear entre duas variáveis, por exemplo, de $X$ e $y$. Se as duas variáveis estiverem simultaneamente acima e abaixo de suas respectivas médias, a co-variância será positiva. Se $X$ estiver acima de sua média quando o $y$ estiver abaixo de sua média ou vice-versa, a co-variância será negativa. O valor da co-variância depende da relação das unidades nas quais $X$ e $y$ forem medidos.

### 4.3.16 - Teste de $x^2$ e Distribuição $x^2$

A distribuição t de Student e em certas ocasiões a distribuição normal podem ser muito úteis para desenvolver testes estatísticos sobre modelos de regressão. A **distribuição** $\boxed{x^2}$ é particularmente útil em testes que envolvem a soma dos quadrados de variáveis distribuídas por meio de uma curva normal.

O $x^2$ começa em uma origem e se distribui para a direita, com uma extremidade que se estende infinitamente para a direita. A formação exata da distribuição depende do número de graus de liberdade uma vez que a distribuição se torna mais simétrica na medida em que um número maior de graus de liberdade for fornecido.

A distribuição $x^2$ pode ser usada para desenvolver e analisar testes estatísticos em uma amostra de variações de uma série de dados obtida a partir de uma população normalmente distribuída.

Através da análise dos valores críticos de $x^2$ com seus graus de liberdade apropriados, e convenientes, o autor decide se é possível ou não rejeitar a hipótese nula de que a variação de séries é igual a um valor positivo específico. Há situações e ocasiões em que o grupo testará as equações simultâneas que envolvem dois ou mais parâmetros de regressão.

Se o grupo quiser testar a hipótese de que a regressão e sua inclinação são ambos iguais a 0 em relação à alternativa de que ou um ou outro é diferente de 0, o teste acima permite este tipo de análise. O teste estatístico adequado baseia-se na função de distribuição F e é caracterizado por dois parâmetros. O primeiro está associado a um número de parâmetros estimados no modelo, o segundo ao número ilimitado de graus de liberdade.

A distribuição F, tal como uma $x^2$ tem uma distribuição assimétrica. É uma distribuição em que a concentração tende mais para a esquerda, e em que posteriormente há uma curva para a direita de inclinação suave que varia de 0 a infinito. Geralmente, o teste estatístico de F envolve a razão entre duas variações independentes.

O valor estatístico F é calculado como a maior estimativa de variação do numerador e da menor estimativa de variação do denominador. Quanto maior for a diferença entre as duas variações, maior será o valor estatístico F. Sendo assim, com um F suficientemente grande ou alto, o autor pode rejeitar a hipótese nula. Na prática, o teste de hipótese é desenvolvido através da escolha de um nível específico de significação, e, posteriormente, pela verificação do valor crítico de distribuição F em um gráfico padrão F.

É igualmente importante assinalar que depois que uma série temporal tenha sido especificada e seus parâmetros estimados, é conveniente efetuar um certo tipo de exame a fim de verificar se a hipótese original era boa ou ruim, se estava correta ou incorreta.

Este processo de diagnóstico geralmente envolve dois momentos, duas etapas. O primeiro consiste na função de autocorreção para as séries simuladas, que pode ser de tal forma que possa ser comparada com a autocorrelação das amostras e com a função de autocorrelação nas séries originais. Caso duas funções de autocorrelação pareçam ser muito diferentes, então isto irá dar origem a uma certa dúvida no que diz respeito à validade do modelo, e será

necessário que esta seja reespecificada e redefinida. No caso das duas funções de autocorrelação não exibirem quaisquer diferenças notáveis, então o teste analítico quantitativo dos resíduos gerados pelo modelo poderá ser feito, e sua credibilidade será maior.

Se o modelo tiver sido corretamente especificado, os resíduos Et, que são as estimativas dos termos de erros não observados, deverão ter as mesmas características e propriedades. Especificamente, deve-se esperar que os resíduos não sejam relacionados uns em relação aos outros, de tal forma que a função de autocorrelação, na amostra de resíduos, seja próxima de zero para qualquer deslocamento de K maior ou igual a 1. Isto significa que uns testes muito eficazes e úteis, calcados em resultados estatísticos nos termos da linha de pensamento de Box podem ser aplicados para uma amostragem de uma função de autocorrelação. Caso o modelo seja correlacionado convenientemente, deslocamentos significativos de K, por exemplo, maiores que 5, podem levar à constatação de que as autocorrelações não estão correlacionadas entre si, com uma média zero e uma variação de 1/t, onde t corresponde ao número de observações nas séries temporais.

Consideremos o valor estatístico R sendo composto do primeiro $K$ residual de correlações de $rl$ a $rk$. Então há $r = t$, a soma de $k = 1$k, para $K$, de . Esta estatística é a soma de variáveis aleatórias normais com quadráticas independentes, cada qual com uma média zero e uma variação 1/t, e estão distribuídas aproximadamente como uma $x^2$. Box mostra que a aproximação é considerável e que o valor estatístico R será distribuído como a $x^2$ , então existirá $x^2$ (K-P-Q) que é a distribuição qui quadrado com K-P-Q graus de liberdade.

**17 - A Distribuição Gama**

Havendo evidências de necessidade, talvez venha a ser trabalhada a função **Distribuição Gama**. O grupo definirá após as primeiras rodadas e testes com dados esta utilização.

A função de densidade da probabilidade da distribuição gama é dada pela fórmula:

$$\mathrm{f}(CX)d\mathrm{x} = \frac{1}{\gamma(\mathrm{n})} \quad e^{-X}X^{n-1}dX, \ \text{com } 0 \le X \le \infty$$

Onde

$$\gamma(n) = \int_0^{\infty} e^{-X} X^{n-1} d\mathrm{x}.$$

representa a função gama.

Naturalmente, ao procurar-se a definição de mensurações ideais, como para o caso de uma amostra de dados que servirá para cálculos de Análise Discriminante Múltipla, um conjunto variado de opções de cálculo estatístico apresenta-se à disposição dos analistas, segundo as suas conveniências e interesses.

Desta forma, caberá a cada analista ponderar, avaliar e meditar sobre os elementos positivos e negativos de cada conjunto de análises possíveis. O correto balanceamento e alinhamento das variáveis com os métodos analíticos poderá contribuir na obtenção de indicadores, dados e informações qualificadas, que ajudarão nos processos decisórios que forem identificados como necessários.