ARNAUD FRANCIS JEAN GUÉRIN

PREVISÃO DO VOLUME DE VENDAS DE UM BEM DE CONSUMO

Trabalho de formatura apresentado
À Escola Politécnica da Universidade de
São Paulo para a obtenção do
Diploma de Engenheiro de Produção

Orientador: Prof. Dr. Álvaro Euzébio Hernandez

São Paulo
2006

FICHA CATALOGRÁFICA

Guérin, Arnaud Francis Jean
Previsão do Volume de Vendas de um Bem de Consumo
p. 116

Trabalho de Formatura – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1. Previsão de vendas 2. Bens de consumo 3. Método analítico quantitativo (previsão) I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

Folha de aprovação

Arnaud Francis Jean Guérin

Previsão do volume de vendas de um bem de consumo.

Trabalho de formatura apresentado
À Escola Politécnica da Universidade de
São Paulo para a obtenção do
Diploma de Engenheiro de Produção

Aprovado em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Prof. Dr. _____

Instituição: _____ Assinatura: _____

Agradecimentos

A Mônica de Cássia Teixeira, pelo apoio de cada instante, pela confiança e colaboração, que foram essenciais ao meu desempenho no Brasil. Mais do que um apoio, um exemplo de integridade e coragem.

Ao Prof. Dr. Álvaro Euzébio Hernandez, pela valiosa orientação durante a execução deste trabalho.

Ao Ronni dos Santos Oliveira pela essencial ajuda lingüística.

A todos na Procter & Gamble que de alguma forma participaram na elaboração deste trabalho.

Resumo

GUÉRIN, Arnaud Francis Jean. **Previsão do volume de vendas de um bem de consumo.** 2006. 116f. Trabalho de conclusão de curso (Trabalho de formatura) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2006.

O propósito deste trabalho de formatura é melhorar a precisão dos métodos de previsão de volume de vendas empregados pela empresa Procter & Gamble. O trabalho se limitará à previsão do volume de vendas de um bem de consumo do mercado de sabão em pó brasileiro. Hoje a empresa Procter & Gamble utiliza modelos de extrapolação para prever o volume de vendas. Primeiramente é feita uma revisão bibliográfica dos métodos de previsão, assim como a determinação daquele mais adequado à luz das características do problema. Como o problema envolve variáveis quantitativas bem conhecidas e que o volume de vendas apresenta variações grandes, verificou-se que a técnica mais adequada é a de regressão linear. Em seguida, comparou-se as precisões obtidas pelos modelos de previsão vigentes (extrapolação) com os modelos propostos (regressão linear). Para isto, adotou-se como critério de comparação o valor do erro padrão. Os modelos de extrapolação apresentam erros padrões em torno de 14%. Já para os modelos de regressão linear, os valores dos erros padrões são próximos e até inferiores a 11%. Portanto foi possível concluir que os modelos de regressão linear são efetivamente mais precisos. Por outro lado, constatou-se que o método de regressão linear tem uma complexidade maior do que os métodos de extrapolação. Desta forma, este fato deve ser levado em consideração no momento da sua escolha, pelo grande investimento em tempo que este implica. Eventualmente, a melhoria da precisão do modelo pode não compensar tantos investimentos.

Palavras-chave: Previsão de vendas. Bens de consumo.

Abstract

GUÉRIN, Arnaud Francis Jean. **Sales´ volume forecasting of a consumer good.** 2006. 116p. Course's conclusion work (Graduation essay) – Polytechnic School, University of São Paulo. São Paulo, 2006.

The purpose of this essay is to improve the precision of sale's volume forecasting models used by the Procter & Gamble Company. This essay will be limited to the sale's volume forecasting of a consumer good in the powder detergent market in Brazil. Procter & Gamble is currently using extrapolation methods to forecast its sale's volume. First, a bibliographical revision of forecasting methods is done, as well as a choice of the most adequate one, based on the problem characteristics. As this problem evolves well known quantitative variables and the sales volume varies a lot, it has been verified that the better technique is the linear regression. After that, the precision obtained by the current forecasting models (extrapolation) are compared with the precision of linear regression models. The criterion of comparison adopted is the value of the standard error. The extrapolation models present standard error around 14%. On the other hand, linear regression models present standard error value close to 11%. So, can be concluded that linear regression method are really more precise. On the other hand, the linear regression method is more complex than the extrapolation ones. The consequence is a bigger time requested by the linear regression method. This large investment must be considered in the decision making process in order to determine if the effort is worth or a waste of energy.

Keywords: Sales forecasting. Consumer goods.

Sumário

1 Introdução

1.1 A empresa

1.1.1 Apresentação geral

A Procter & Gamble (P&G) é uma empresa multinacional de bens de consumo fundada em 1837, em Cincinnati, Ohio – Estados Unidos. Atualmente, a P&G comercializa aproximadamente 300 marcas, em mais de 160 países, operando em cerca de 80 países.

A atuação, no Brasil, se iniciou em 1988, com a aquisição da empresa Perfumarias Phebo S.A. e hoje conta com cerca de 1600 funcionários, faturando 432 milhões de dólares. A P&G possui duas fábricas no estado de São Paulo: Anchieta, que produz sabão em pó; e em Louveira, produz o restante dos produtos (Pantene, Pampers, Always etc)

No fim de 2005, a aquisição da empresa The Gillette Company, proporcionou à P&G uma maior atuação no mercado de bens de consumo, focando também o público masculino. The Gillette Company possui uma única fábrica localizada em Manaus.

1.1.2 Visão e estratégia da empresa

A visão da empresa é “Ser, e ser reconhecida como a melhor companhia de produtos de bens de consumo do mundo”. Visando alcançar esta meta, a P&G elaborou uma estratégia de crescimento baseada em dois pontos:

- Onde Atuar:

1. Tornar os negócios principais em líderes globais.

2. Fazer crescer grandes marcas, mercados e clientes.

3. Desenvolver negócios que crescem mais rapidamente, com maior margem de lucro.

4. Restabelecer a liderança na Europa Ocidental.

5. Acelerar o crescimento nos mercados de baixo poder aquisitivo.

- Como Vencer:

1. O consumidor é o chefe.

2. Vencer no 1º e 2º momentos da verdade (compra e uso, respectivamente).

3. Entregar o melhor custo, fluxo de caixa e produtividade.

4. Alavancar excelência organizacional e operacional.

1.1.3 Os produtos da empresa

Este trabalho irá focar, exclusivamente, os produtos da P&G, não expondo os produtos da Gillette. A razão desta escolha se deve ao fato de que, no momento da redação, a integração entre as duas companhias não tinha sido concluída, ou seja, estas funcionavam de forma independente.

A estratégia da companhia em não associar diretamente o seu nome com suas marcas, resulta em que os nomes das marcas sejam mais famosas no mercado do que o nome da empresa P&G. Hoje, a P&G se baseia sobre 13 marcas globais fortes que faturam mais de um bilhão de dólares, entre elas Pampers, Always, Ariel, Crest etc.

No Brasil, a P&G comercializa produtos dentro de seis categorias:

- Cuidados com o Lar – os sabões em pó (Ariel, Ace, Bold e Pop).
- Cuidados com o Bebê – as fraldas descartáveis (Pampers).
- Proteção Feminina – os absorventes (Always e Tampax).

- Cuidados com a Beleza – os shampoos, condicionadores (Pantene) e colorantes para cabelo (Wella Color).
- Cuidados com a Saúde – os remédios (Hipoglós), pastas de dente (Crest), etc.
- Alimentos – os salgados (Pringles).

A maior parte dos produtos referentes às duas últimas categorias citadas acima é importada de fábricas da P&G fora do Brasil. Vale também ressaltar que Hipoglós é uma marca exclusivamente brasileira, não existente em outros pais.

1.2 O estágio

O estágio foi desenvolvido na área de finanças da P&G. De maneira simplificada, a área de finanças pode ser dividida em três grupos.

1-) Gerenciamento do setor de impostos, contabilidade etc.

2-) Analistas financeiros: responsáveis pelo controle das categorias, baseando-se no desempenho de cada categoria de produto.

3-) Atuação na área de vendas (na qual realizo estágio), fornecendo apoio às equipes multifuncionais de vendas visando a melhora da eficiência dos investimentos.

1.2.1 Estrutura da área de vendas

Será dado enfoque à estrutura da área de vendas para que se entenda a necessidade de um trabalho multidisciplinar nesta área.

A área de vendas é dividida em equipes multifuncionais. Cada equipe atende um tipo de cliente, desenvolvendo com ele uma relação particular, propiciando, assim, um ambiente de parcerias e trabalho em conjunto; realidade que se torna ainda mais visível para os maiores clientes, que possuem uma equipe exclusivamente dedicada a eles.

Uma equipe multifuncional é constituída de: um líder, representantes de vendas que são apoiados por profissionais chamados multifuncionais das áreas de logística, finanças, marketing e sistemas. Em cada equipe de vendas há, pelo menos, um multifuncional de cada área, sendo que o número de profissionais poderá variar de acordo com a característica de cada equipe.

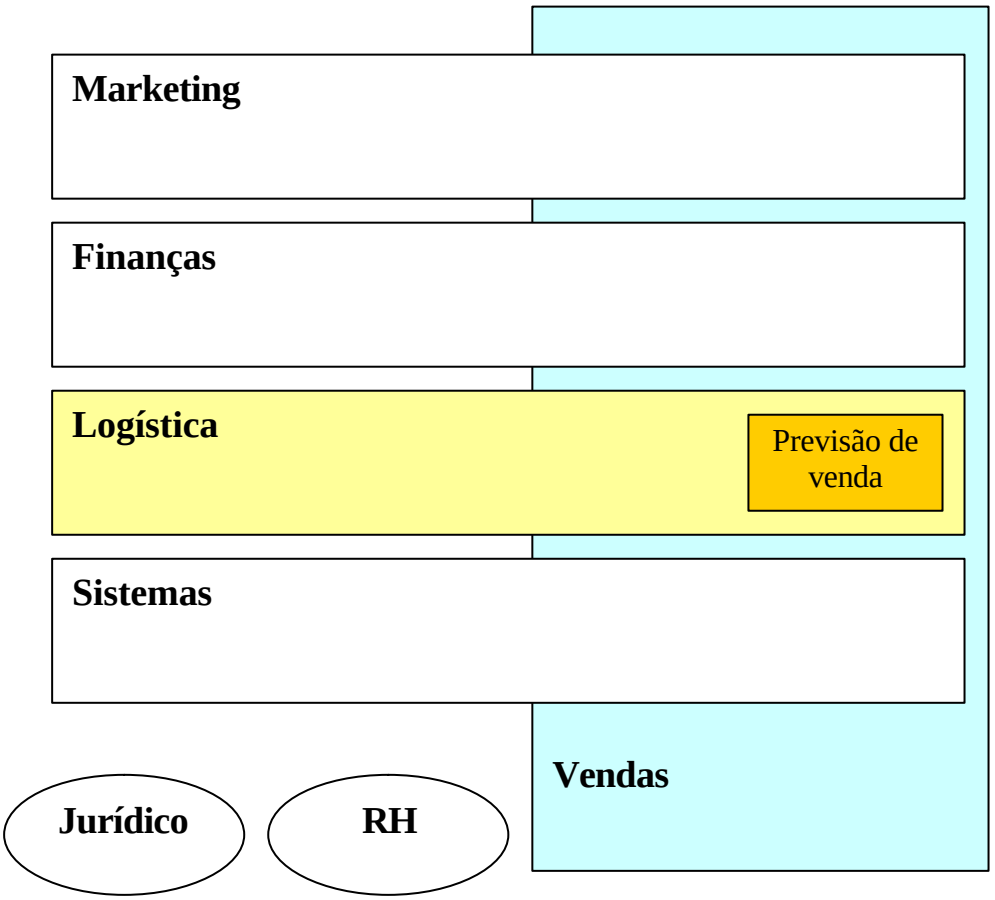
Assim, pela diversidade das missões e dos contatos entre diferentes áreas, a cultura e o aprendizado de um engenheiro de produção são úteis para uma visão geral necessária à resolução dos problemas nas equipes multifuncionais.

1.2.2 O dia-dia de estagiário

Como já citado anteriormente, estou desenvolvendo o estágio dentro de uma equipe multifuncional de vendas, atuando como integrante da área de finanças. A minha tarefa é acompanhanhamento e análise do desempenho dos produtos da companhia para os clientes da minha equipe, através de relatórios mensais e de elaboração de Scorecards, assim como controle do budget da equipe. No decorrer do mês, acrescentam-se atividades adicionais que correspondem a projetos específicos como, por exemplo, análise financeira do impacto de uma promoção para um cliente.

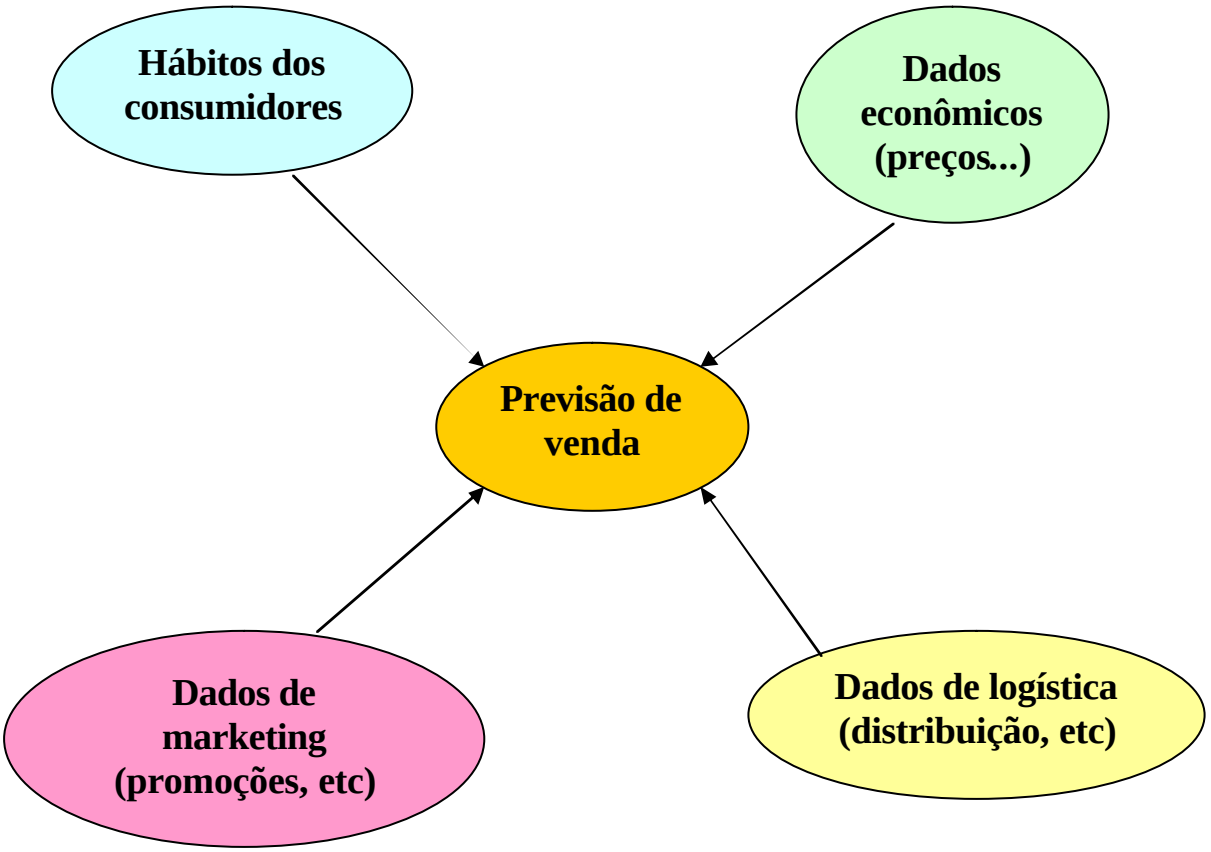
1.3 Apresentação do problema

O ramo da previsão é desafiador e se tornou essencial para as empresas tomarem grandes decisões estratégicas ou gerenciarem o seu negócio. Assim, nos últimos anos, a P&G decidiu investir recursos para alcançar um melhor nível de precisão de suas previsões de volume de vendas. A área de previsão de vendas se situa no departamento de logística voltado a venda, como organograma 1 a seguir o ilustra:



Organograma 1 – Departamentos da empresa

Os modelos atuais usados para prever os volumes de vendas dos produtos pertencem todos à categoria dos modelos de série temporal. Os planejadores possuem uma ferramenta própria no sistema Enterprise Ressource Planning (ERP) para elaborar as previsões. Os tipos de modelos disponíveis são apresentados no segundo capítulo: média móvel, suavização com tendência, suavização com sazonalidade etc. Os planejadores, inicialmente, estudam os dados da variável a prever para determinar suas características para, assim, escolher o modelo de série temporal mais adequado. Uma vez que a previsão é feita com o modelo, os planejadores trabalham nesse resultado para aperfeiçoá-lo, levando em conta as iniciativas promocionais durante o período, o preço do produto e da concorrência, além de todas as variáveis que impactam no volume de vendas. O esquema 1 a seguir ilustra os grandes tipos de variáveis que influenciam no volume de vendas:



Esquema 1 – Tipos de variáveis influenciando a venda

Por fim, o objetivo dos planejadores é chegar a um nível de precisão ainda maior na previsão dos volumes dos principais produtos vendidos pela companhia com o propósito de tomada de decisões que permitem o sucesso da empresa a médio e longo prazos. A empresa gostaria de investigar a possibilidade de aperfeiçoar seus métodos de previsão atualmente empregados. O trabalho aqui desenvolvido tratará de **Previsão do volume de vendas de um bem de consumo**, pois estudará a possibilidade de melhoria dos métodos de previsão da empresa, no caso especifico de um produto por ela vendido. Contudo, neste trabalho, restringe-se ao mercado de sabão em pó.

1.4 Quadro de referência do trabalho

Uma vez o problema definido, vale a pena se perguntar se ele entra no quadro de estudo da engenharia de produção: quais são os critérios da engenharia de produção a que ele corresponde?

De acordo com a definição clássica, adotada tanto pelo American Institute of Industrial Engineering (A.I.I.E.) como pela Associação Brasileira de Engenharia de Produção (ABEPRO),

> "Compete à Engenharia de Produção o projeto, a implantação, a melhoria e a manutenção de sistemas produtivos integrados, envolvendo homens, materiais e equipamentos, especificar, prever e avaliar os resultados obtidos destes sistemas, recorrendo a conhecimentos especializados da matemática, física, ciências sociais, conjuntamente com os princípios e métodos de análise e projeto da engenharia."

Essa definição frisa o caráter multidisciplinar da engenharia de produção, deixando ambígua a fronteira com outras disciplinas como a administração, por exemplo. Para especificar mais o que é engenharia de produção, o departamento de engenharia de produção da Universidade

Federal de Minas Gerais explica que toda engenharia é: *"uma ciência aplicada", cujos problemas são resolvidos recorrendo-se aos conhecimentos de ciências "puras", das ciências sociais e aos métodos da engenharia*".

Assim, o tema abordado nesse trabalho: **Previsão do volume de vendas de um bem de consumo** se enquadra à descrição acima. O trabalho se trata-se da aplicação de métodos matemáticos com todo rigor de um engenheiro, para se chegar a um resultado prático, manipulado facilmente pelo usuário: um ou vários modelos de previsão. Essa praticidade no uso vem do fato de que o modelo é uma representação simplificada da realidade para uma finalidade. A primeira finalidade seria prever as vendas de um bem de consumo, conhecendo antecipadamente seu histórico. A partir deste ponto, levanta-se uma pergunta bem legitima: por que prever? O capitulo a seguir, de revisão bibliográfica, fornecerá alguns elementos de resposta para essa pergunta.

1.5 O nosso caminho

Após a introdução e definição do problema colocado pela empresa, é feita a revisão bibliográfica dos pontos teóricos necessários ao entendimento do trabalho, bem como a determinação dos métodos potencialmente adequados para resolver o problema. Primeiramente são definidos e detalhados alguns conceitos ligados à previsão. Depois são apresentadas as principais ferramentas necessárias às analises dos dados a prever, a definição dos critérios de medição de erro utilizados para avaliar os modelos, os principais modelos utilizados até a presente data pela companhia. Na finalização do capitulo dois, apresenta-se a teoria simplificada do principal modelo explicativo, levando em conta dentro do seu escopo, as variáveis que tem impacto no volume a prever.

No terceiro capítulo, serão seguidas as etapas do modelo explicativo: o método de regressão multilinear, selecionando as combinações de variáveis importantes, testando-as, e, por fim, validando as configurações de melhor desempenho. Os resultados dos modelos de extrapolação também serão apresentados.

Para validação deste trabalho, no capitulo quatro, serão confrontados os melhores modelos explicativos aos modelos de extrapolação utilizados pela empresa, para assim se verificar a validação da melhoria da previsão apresentada.

No quinto capítulo, haverá o resumo dos principais pontos de aprendizados deste trabalho, com as conclusões alcançadas, e, por fim, fornecimento de dicas para o aprofundamento do estudo do problema.

Finalmente, os dois últimos capítulos compreender-se-ão em referências de apoio, e apêndices.

2 Revisão Bibliográfica e Métodos empregados

Primeiramente, serão apresentadas as razões que motivam as empresas a investir no ramo da previsão e os ganhos ao investir em previsão de volume de vendas. Posteriormente, será explicada, de maneira adequada, a teoria matemática necessária à resolução do problema. Serão citadas todas as grandes famílias de métodos de previsão neste capitulo, assim como as razões que levam a usar tal ou tal tipo de modelo. Por fim, serão analisadas, com mais detalhes, as técnicas relevantes ao nosso problema de **previsão do volume de vendas de um bem de consumo.**

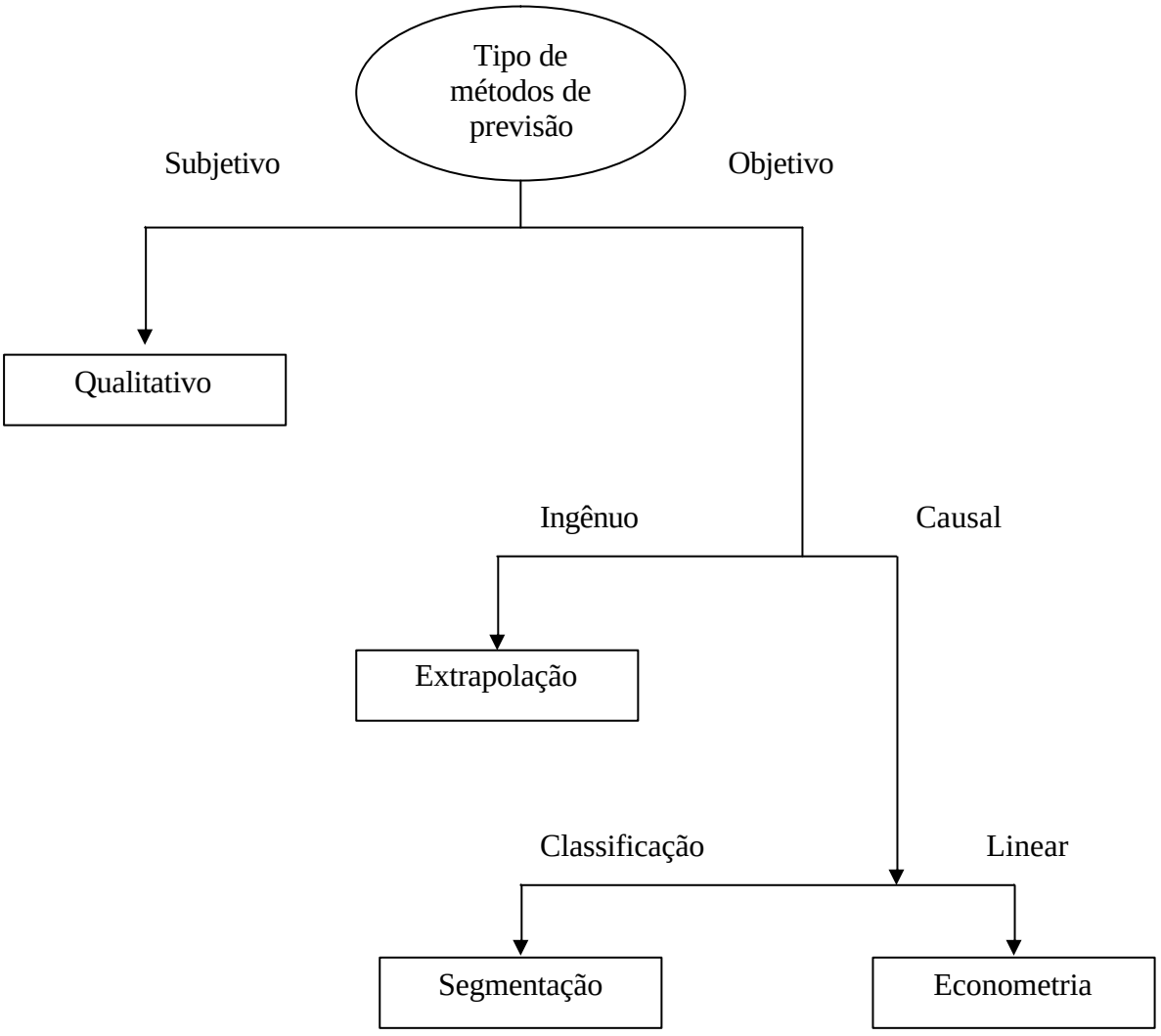
2.1 A necessidade de previsões

Um dos fatores críticos de sucesso, no mundo empresarial, é o conhecimento do seu ambiente e das variáveis que impactam o seu negócio. A maioria das empresas conhece bem o seu mercado e seus “atores”: concorrentes, consumidores etc. Segundo Porter, cada empresa deve ter uma estratégia clara para assegurar a continuidade do seu sucesso. A noção de estratégia envolve elementos do futuro não conhecidos. Desta forma, as empresas investem no ramo de previsão para obter o máximo de informações que serão base para a tomada de decisões importantes. Assim, quanto mais precisa a previsão, menor será o risco para uma determinada decisão.

Para previsão de volume de vendas de um produto, quanto mais preciso o volume é estimado, mais adequadas serão as quantidades de matéria prima a serem compradas e de produtos a serem fabricados, melhor serão utilizados os centros de distribuição etc. Estas melhorias resultam em economias que asseguram o futuro da empresa.

2.2 Métodos de previsão

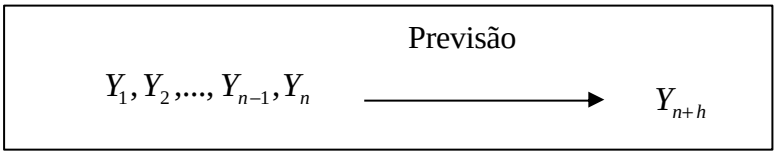
Segundo Armstrong (1985), pode-se estruturar os diversos métodos de previsão com a ajuda dessa árvore (esquema 2), a seguir:



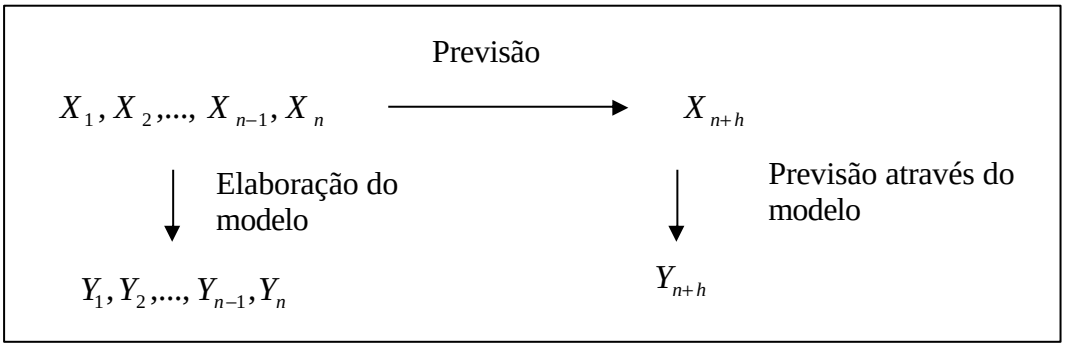
Esquema 2 – Árvore dos métodos de previsão.
Adaptado de Armstrong, (1985).

No ramo da previsão existem quatro grandes tipos de técnicas, como visto na árvore precedente. Usar um tipo ou um outro depende das decisões do planejador frente às

informações que ele tem. Para elaborar um modelo quantitativo como no nosso caso, Makridakis; Wheelwright; Hyndman (1998) explicam que nós devemos seguir o caminho dos modelos objetivos na árvore. Segundo eles, os modelos subjetivos devem ser escolhidos quando pouca ou nenhuma informação quantitativa está disponível. No caso desse trabalho, tem-se acesso a varias medidas quantitativas. Assim, o trabalho orienta-se do lado dos métodos objetivos. Depois disso, nós temos uma segunda escolha a fazer entre métodos ingênuos (em inglês *naive*) e causais. Para isso, precisa-se explicar quais são as diferenças entre esses métodos. Os esquemas 3 e 4, a seguir nós ajudarão nesta tarefa:



Esquema 3 – Principio do método ingênuo



Esquema 4 – Principio do método causal

Com X_i as variáveis independentes,

Y a variável dependente, que sofre a previsão,

n o tamanho da amostra,

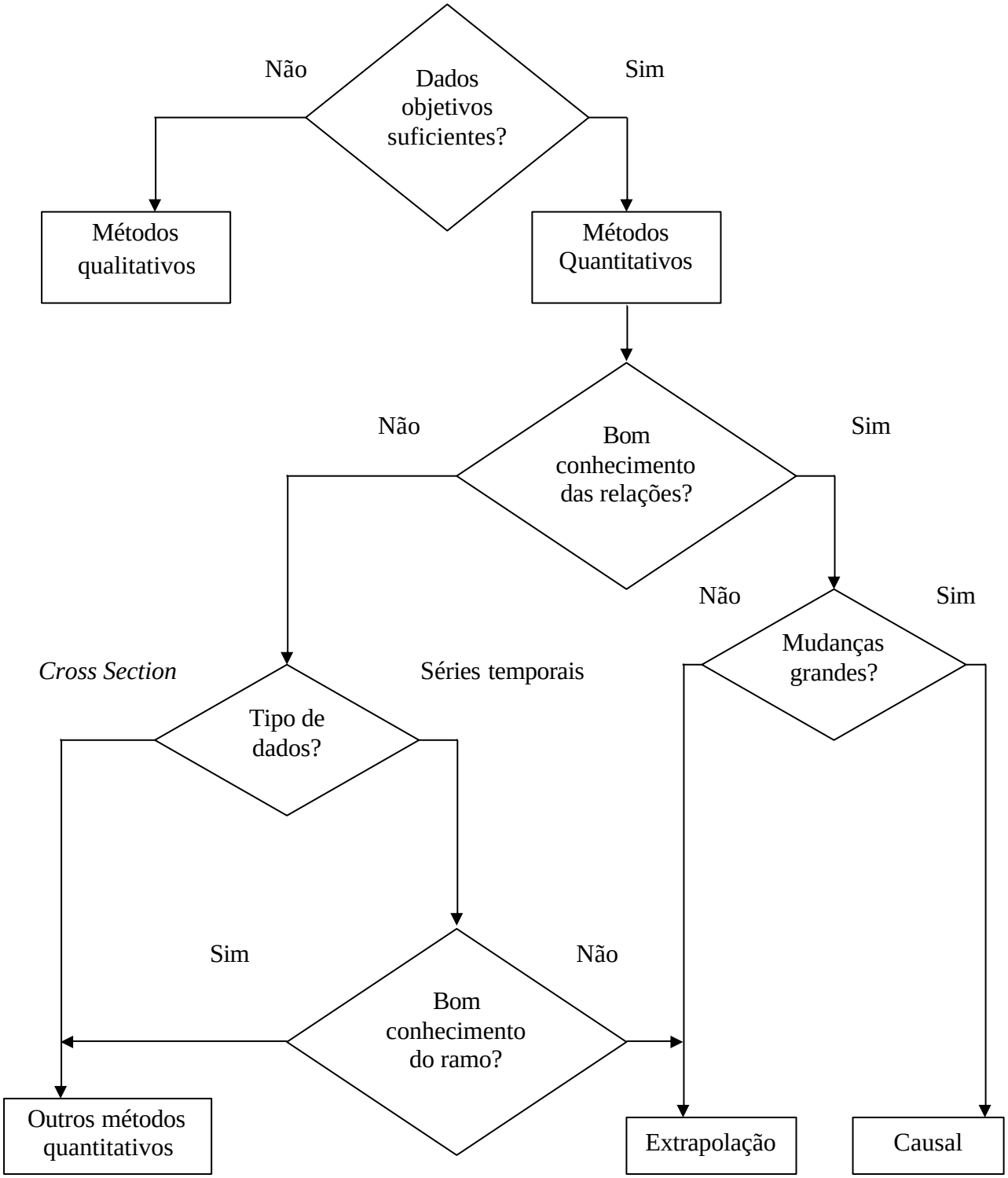
h a posição futura do período a prever no futuro.

Pode-se ver que o método ingênuo baseia-se no histórico da variável estudada para fazer sua projeção. De certa maneira, o método olha o passado da variável para elaborar uma projeção dela no futuro.

O método causal segue um caminho diferente em várias etapas. Primeiro estudam-se as relações entre as variáveis independentes X_i e a variável dependente Y a prever com bases os históricos das variáveis respectivas. Depois disso, vão ser previstas as variáveis independentes para o período estipulado para usar o modelo e assim prever a variável dependente. Armstrong (1985) explica que os modelos causais têm um poder explicativo que pode ser bem interessante, principalmente para previsão de uma variável que varia muito. Para que o método seja eficiente, as variáveis independentes, que vão explicar a variável estudada, precisam ser simples a prever.

2.3 Determinação dos métodos a serem empregados

A seguir, no fluxograma 1, está apresentado o raciocínio simplificado de Armstrong, (2001), para a escolha do método dentre os quatros disponíveis:



Fluxograma 1 – Árvore de decisão do método de previsão
Adaptado de Armstrong, (2001).

No caso deste trabalho, partindo do início da árvore precedente, existem vários dados quantitativos / objetivos para medir o desempenho dos produtos do mercado de sabão em pó. Portanto, orienta-se na direita da árvore de decisão. O passo seguinte é se perguntar se existe um bom conhecimento das relações entre os dados objetivos. Um exemplo simples seria a relação entre o preço do produto P&G e as suas vendas. Se o preço aumenta, a tendência das vendas é de diminuir. Podem ser observadas relações simples assim com a maioria das variáveis em nossa possa. De novo, orienta-se na direita na árvore para se perguntar se a variável estudada é sujeita a grandes variações. Pode se observar que no período estudado o volume de vendas (variável estudada) varia entre um valor mínimo e um valor maximo igual mais ou menos a quatro vezes o valor mínimo. Esse ponto demostra as grandes variações do volume de vendas do produto P&G. Pode-se observar também que com certas variáveis como o número de promoções da concorrência, mudanças muito grandes de volume de vendas do produto P&G ocorrem. Estes fatos levam a escolher o tipo de método causal.

Por fim, segundo Armstrong (1985), a última escolha entre os métodos lineares e de classificação (na primeira árvore) é a de menor importância e na maioria das vezes, são privilegiados modelos econométricos lineares. Assim, são apresentados os modelos econométricos lineares que parecem relevantes para buscar uma solução do problema. Também, são explicados os modelos de extrapolação usados hoje na empresa para que se possa fazer uma comparação do desempenho desses dois tipos de modelos.

2.4 Detalhamento dos métodos

Os parágrafos 2.4.1 e 2.4.2 são dedicados a conceitos válidos para os dois métodos quantitativos explicados a seguir. O primeiro enfoca-se na caracterização dos dados, quando o segundo apresenta medidas de acurácia.

2.4.1 Características dos dados

Ao iniciar a elaboração de um modelo de previsão, é muito importante estudar as características dos dados disponíveis. Com a primeira etapa de caracterização dos dados, os planejadores escolherão o tipo de modelo mais adequado para elaboração de uma previsão precisa.

2.4.1.1 Uma variável

As medidas descritas a seguir são as mais utilizadas para descrever dados de uma série.

Primeiramente, existem duas maneiras de definir o “centro” de uma série de dados:

- A mediana de uma série de N valores é a medida $\frac{N}{2}$ quando as medidas são classificadas por ordem crescente.
- A média é definida da seguinte maneira para uma série de N dados $\{X_1,..., X_N\}$:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad \textbf{(2.1)}$$

Uma idéia essencial quando se descreve uma série de dados é a variabilidade. Essa idéia até parece no momento da escolha do tipo de modelos a serem usados. De maneira geral, esta variabilidade é medida em relação à média. Assim, o desvio da média é definido da seguinte forma:

$$\left(X_i - \overline{X}\right) \tag{2.2}$$

A soma destes desvios sempre será igual a zero, assim, precisam-se de medições úteis para caracterizar a variabilidade da série de dados.

A primeira dela é o *Mean of the Absolute Deviations (MAD)* definido da seguinte maneira:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \overline{X}\right| \tag{2.3}$$

Uma outra medida é o *Mean of Square Deviations (MSD)* caracterizada por ponderar de forma mais forte os desvios de média maior:

$$MSD = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \tag{2.4}$$

Uma medida bem próxima desta é a Variância S^2: definida como soma dos desvios da média dividida pelo número de graus de liberdade. O número de grau de liberdade é definido como o número de dados menos o número de parâmetros estimados. Como a média é

estimada, o número de grau de liberdade é N-1. A variância é definida a seguir na equação 2.5:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \qquad \textbf{(2.5)}$$

A partir da variância, define-se o desvio padrão S (ou *Standard Deviation*) que é a raiz quadrada da variância (equação 2.6). Uma propriedade importante do desvio padrão é ter a mesma unidade do que os dados.

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \qquad \textbf{(2.6)}$$

2.4.1.2 Duas variáveis

Sejam $\{X_1,...,X_N\}$ e $\{Y_1,...,Y_N\}$ duas séries de dados estudadas. Uma medida importante no caso de duas variáveis é a Covariância que permite estudar as relações de comportamento entra as duas variáveis e o quanto elas variam juntas. A Covariância é definida da seguinte maneira:

$$Cov_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \qquad \textbf{(2.7)}$$

O problema da Covariância é a dificuldade de interpretação devido à unidade desta medida. Para resolver este problema usa-se o Coeficiente de Correlação definido a seguir (2.8):

$$r_{XY} = \frac{Cov_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}} \qquad \textbf{(2.8)}$$

A Covariância e o Coeficiente de Correlação são medidas para quantificar uma relação linear entre duas variáveis diferentes. Já a Autocovariância e a Autocorrelação têm o mesmo objetivo para uma única série de dados. A relação linear será quantificada entre dados da mesma série.

A Autocovariância c_k, onde k é a defasagem entre os intervalos de dados estudados, é definida da seguinte maneira:

$$c_k = \frac{1}{N} \sum_{i=k+1}^{N} \left(X_i - \overline{X}\right)\left(X_{i-k} - \overline{X}\right) \qquad \textbf{(2.9)}$$

A Autocorrelação r_k é definida desta maneira:

$$r_k = \frac{\sum_{i=k+1}^{N} \left(X_i - \overline{X}\right)\left(X_{i-k} - \overline{X}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \qquad \textbf{(2.10)}$$

Estas duas medidas (2.9 e 2.10) são bem úteis para se detectar relação temporal de causa e efeito dentro de uma mesma série de dados.

2.4.2 Medida da acurácia da previsão

Conhecendo as características dos dados a estudar, é muito mais fácil determinar quais serão os modelos mais adequados para a previsão. Uma pergunta surge: como saber se um modelo é melhor do que outro? Neste parágrafo, estudam-se as principais formas de se medir o erro de previsão de um modelo. Quando se comparam dois métodos de medição de acurácia em um mesmo modelo de previsão, geralmente, os resultados não são muito diferentes. Mas, o modelo pode ser mais preciso com um método do que com outro, o que ressalta a importância de entender o funcionamento de cada medição de acurácia.

O erro de previsão é definido como a diferença entre o valor real Y_t e o valor estimado na mesma data F_t:

$$e_t = Y_t - F_t \qquad \textbf{(2.11)}$$

Existem 3 parâmetros principais para medir a acurácia de um modelo:

- O *Mean Error (ME)*: permite medir a presença e a direção de um viés. Quando é positivo, a evolução dos valores é superestimada. A definição do *ME* é:

$$ME = \frac{1}{N}\sum_{t=1}^{N} e_t \qquad \textbf{(2.12)}$$

O problema desta medida é que os erros têm valores algébricos e quando somados se anulam um com o outro.

- O *Mean Absolute Error (MAE)* ou *Mean Absolute Deviation (MAD)*: permite examinar o tamanho dos erros de previsão. Sua definição é (com as mesmas convenções):

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|e_t\right| \qquad \textbf{(2.13)}$$

Na resolução do problema deste trabalho será utilizada uma medida relativa do MAE: o erro padrão que é definido como o MAE dividido pela média dos valores:

$$Erro\,Padrão = \frac{MAE}{\overline{X}} \qquad \textbf{(2.14)}$$

- O *Root Mean Square Error (RMSE):* meio alternativo para examinar o tamanho dos erros de previsão. Neste caso, os erros maiores terão pesos bem maiores devido ao quadrado. Uma vantagem é que a unidade desta medida também é a mesma que os dados. Sua definição é (com as mesmas convenções):

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}e_t^{\,2}} \qquad \textbf{(2.15)}$$

2.4.3 Modelos de extrapolação

Como explicado anteriormente na seção 2.3, são dois os modelos relevantes para o presente problema: os modelos de série temporais e os modelos de regressão multilineares.

Nesta parte, apresentam-se os modelos de série temporal, ou extrapolação, usados atualmente pela empresa P&G. Na parte a seguir, descreve-se o modelo explicativo de regressão multilinear.

Uma série temporal é um conjunto de medidas observadas de uma quantidade, ao longo do tempo, onde os intervalos de tempo são constantes.

De modo geral, é assumido pela literatura que uma série temporal possui três componentes de pesos variáveis que a compõem:

- Um termo de sazonalidade
- Um termo de tendência/ciclicidade
- Um termo irregular

A representação matemática desta idéia pode ser feita da seguinte maneira (2.16):

$$x_t = f(s_t, t_t, i_t) \qquad \textbf{(2.16)}$$

Com x_t o valor da série temporal no instante t

s_t o componente sazonal no instante t

t_t o componente de tendência/ciclicidade no instante t

i_t o componente irregular no instante t

Segundo Pindyck, Rubinfeld, (1998), a decisão de escolher um modelo de série temporal ocorre quando pouca ou nenhuma informação é conhecida sobre os fatores que impactam a variável estudada, quando um grande número de dados do passado são

disponíveis e quando se quer prever a evolução da variável estudada no curto prazo. Os principais métodos são apresentados a seguir.

2.4.3.1 Média móvel

Sejam $\{x_1 ... x_t\}$ valores observados de uma série temporal durante o período t. O método de previsão da média móvel vai usar estas observações para prever o valor do instante t+1: $f_{t,1}$. O t indica o instante no qual se está e o 1 indica para qual próximo passo se estará prevendo: o instante t+1. A definição matemática de $f_{t,1}$ é apresentada em 2.17 a seguir:

$$\begin{aligned} f_{t,1} &= \text{média das N últimas observações} \\ &= \frac{1}{N}\sum_{i=1}^{N} x_{t+1-i} \end{aligned} \qquad \textbf{(2.17)}$$

N é um inteiro dado que deve ser escolhido de forma a minimizar o *Mean Absolute Error (MAE)* definido anteriormente. Neste caso, o e_t é definido da seguinte maneira: $e_t = x_t - f_t$.

2.4.3.2 Suavização exponencial simples

Este método simples e sem necessidade de cálculos complexos é usado na presença de uma série temporal que flutua em torno de um nível base.

Define-se A_t como a previsão feita no instante t observando o valor x_t, para o instante seguinte t+1:

$$A_t = \alpha x_t + (1-\alpha)A_{t-1} \quad \textbf{(2.18)}$$

O fator α é a constante de suavização que satisfaz $0<\alpha<1$. A previsão A_t é uma média ponderada pela constante de suavização do último valor observado x_t com a previsão para o último valor observado A_{t-1}. Com um valor alto de α dá-se um peso maior para a última observação x_t.

O termo de erro no instante t pode ser definido como $e_t = x_t - A_{t-1}$. Assim, pode-se escrever a definição de A_t de uma outra forma:

$$A_t = A_{t-1} + \alpha e_t \quad \textbf{(2.19)}$$

Com esta nova maneira de escrever (2.19), constata-se que a previsão para o instante t+1 é a previsão para o instante t corrigido de uma fração do erro que foi feito para prever o instante t. Assim se superestimasse x_t, diminuí-se a previsão, no caso inverso, aumenta-se a previsão. O valor de α adequado é o que minimiza o *MAE*.

2.4.3.3 Suavização exponencial com tendência: Método de Holt

O Método de Holt apresenta boas previsões com séries temporais com tendência, mas sem sazonalidade.

Este método é baseado sobre a previsão de duas variáveis compondo a série temporal: o nível de base L_t e a tendência por período T_t. A cada uma das variáveis está alocada uma

constante de suavização, a para L_t e b para T_t, para corrigir as evoluções de nível de base L_t e da tendência por período T_t. Estas constantes respondem nas seguintes regras: 0< a <1 e 0< b <1.

Após observar o dado do instante t: x_t, as equações a seguir permitem estimar os valores de nível de base e de tendência para estabelecer a previsão:

$$L_t = a x_t + (1 - a)(L_{t-1} + T_{t-1}) \quad \textbf{(2.20)}$$

$$T_t = b(L_t - L_{t-1}) + (1 - b)T_{t-1} \quad \textbf{(2.21)}$$

Para calcular a previsão do nível base L_t, usa-se uma média ponderada de x_t que é o último valor observado, com $(L_{t-1} + T_{t-1})$ que é a previsão do nível base do instante t.

Para calcular a previsão da tendência por período T_t, usa-se uma média ponderada da estimativa de aumento de nível base $(L_t - L_{t-1})$, com a última previsão da tendência.

Assim a previsão para o dado x_{t+k} feita no instante t será $f_{t,k}$ com:

$$f_{t,k} = L_t + kT_t \quad \textbf{(2.22)}$$

No caso particular da previsão no instante t para o instante t+1, a previsão será:

$$f_{t,1} = L_t + T_t \quad \textbf{(2.23)}$$

2.4.3.4 Suavização exponencial com sazonalidade: Método de Winter

O método de Winter é usado para fazer previsões com série de dados com sazonalidade e tendência.

Como no método de Holt, separa-se o nível base da tendência por período. Neste modelo aparece uma terceira variável a prever, a variável s_t que é uma estimativa do fator multiplicativo sazonal no instante t. A constante c será o período desta sazonalidade.

As três variáveis a prever, para obter uma previsão na série de dados estudados, são definidas a seguir:

$$L_t = a\,\frac{x_t}{s_{t-c}} + (1-a)(L_{t-1} + T_{t-1}) \qquad \textbf{(2.24)}$$

$$T_t = b\,(L_t - L_{t-1}) + (1-b)\,T_{t-1} \qquad \textbf{(2.25)}$$

$$s_t = g\,\frac{x_t}{L_t} + (1-g)\,s_{t-c} \qquad \textbf{(2.26)}$$

Observa-se que no cálculo do nível base L_t, tira-se a sazonalidade de x_t dividindo este pelo fator de sazonalidade de um período c atrás s_{t-c}, para realmente achar o nível base sem variação devida à sazonalidade. As oscilações devidas à sazonalidade são incluídas na variável s_t.

Para calcular o fator de sazonalidade, usa-se uma média ponderada do mais recente fator adequado para o período s_{t-c}, com uma estimativa da sazonalidade no instante t+1 $\frac{x_t}{L_t}$.

Assim, a previsão para o dado x_{t+k} feita no instante t será $f_{t,k}$ com:

$$f_{t,k} = (L_t + kT_t)\,s_{t+k-c} \qquad \textbf{(2.27)}$$

No caso particular da previsão no instante t para o instante t+1, a previsão séria:

$$f_{t,1} = (L_t + T_t)s_{t+1-c} \qquad \textbf{(2.28)}$$

2.4.4 Regressão linear

Após ter explicado sucintamente os métodos de extrapolação, é detalhado o principal método causal que ajudará a resolver o problema: o método de regressão linear.

Muitas vezes na literatura nenhuma distinção é feita entre econometria e regressão linear. Segundo Jarrett (1987), econometria significa medida econômica, mas nem todas medidas econômicas pertencem ao ramo da econometria. A econometria é uma disciplina que tenta estabelecer relações entre variáveis econômicas graças a teoria estatística. A regressão linear é um importante caso particular da econometria.

Uma variável dependente é o que se quer descobrir, prever. As variáveis independentes são as que possuem um tipo de influência sobre a variável dependente.

O objetivo de uma regressão linear é achar a relação entre uma variável dependente e as variáveis independentes das quais esta depende, para assim, conhecendo os valores destas variáveis independentes, ter acesso a uma estimativa da variável dependente.

O método de regressão linear se baseia sobre as variáveis explicativas (variáveis independente) que têm impacto sobre a variável estudada (variável dependente). Esta abordagem é muito diferente daquela feita com os modelos de previsão de séries temporais que se baseiam unicamente sobre o histórico da variável estudada para prever a sua evolução futura.

Ao escolher um método de regressão linear, é importante saber que o conhecimento do sistema estudado deve ser amplo, e que recursos muito maiores do que nos modelos de séries temporais serão gastos para entender o ambiente no qual a variável dependente esta evoluindo, quais as relações com o seu ambiente etc. Isto significa, na maioria das vezes, um investimento maior em dinheiro e energia. Assim, é importante uma avaliação prévia para adequar o modelo com as expectativas e necessidades: o ganho de precisão do modelo de regressão linear nem sempre cobre o investimento em tempo e recursos que este necessita.

Antes de expor os principais pontos teóricos necessários para elaborar um modelo de regressão linear, estuda-se o caso mais simples deste: o modelo de regressão linear simples de uma variável. Esta etapa preliminar é importante para se obter um pouco mais de intuito e sensibilidade com a teoria, a seguir, do modelo multilinear de k variáveis.

2.4.4.1 Regressão linear simples

O objetivo da regressão simples é conhecer as relações entre a variável dependente y e uma variável independente x, o quanto elas se afetam, como se comportam as variáveis. Neste caso, assume-se que todas as outras variáveis estão constantes.

A relação matemática procurada entre x e y é:

$$y_i = a + bx_i + e_i \qquad \textbf{(2.29)}$$

Com e_i o erro entre o valor medido y_i e $\hat{y}_i = \hat{a} + \hat{b}x_i$ o valor estimado, no instante i.

Os valores $\hat{a}$ e $\hat{b}$ dos estimadores de a e b são escolhidos de acordo com o método dos Mínimos Quadrados. Este método estipula que os coeficientes a e b, para a reta aproximar-se da melhor maneira os valores reais, devem minimizar a seguinte expressão:

Assim, pode-se deduzir $F(\hat{a},\hat{b}) = \sum_{i=1}^{N} \hat{e}_i^2 = \sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2$ por derivações da expressão precedente ($\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$) que: **(2.30)**

$$\hat{b} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad \textbf{(2.31)}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad \textbf{(2.32)}$$

Certamente este modelo é muito simplista e não oferece na maioria do tempo a precisão exigida pelo planejador.

Agora definir-se-ão as principais noções teóricas necessárias para entender o modelo de regressão linear múltipla assim como as suas hipóteses.

2.4.4.2 Regressão linear múltipla

2.4.4.2.1 Modelo de regressão linear de k variáveis.

O caso mais geral da regressão linear múltipla de k-1 variáveis independentes é dado pela equação a seguir:

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + ... + b_k X_{ki} + e_i \quad \textbf{(2.33)}$$

Com Y variável dependente

X ´s variáveis independentes

i índice da observação, variando de 1 a n com n número de observações:

X_{1i} seria a i-ésima observação da variável independente X_1.

e_i termo de erro na i-ésima observação.

b_1 intercepto

b_2 a b_k coeficientes de inclinação

Devido à complexidade da notação, fica mais simples de apresentar esta mesma equação de forma matricial:

$$Y = Xb + e \quad \textbf{(2.34)}$$

Com

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ : \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{21} & .. & X_{k1} \\ 1 & X_{22} & .. & X_{k2} \\ : & : & : & : \\ 1 & X_{2n} & .. & X_{kn} \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ : \\ b_k \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ : \\ e_n \end{bmatrix} \quad \textbf{(2.35)}$$

2.4.4.2.2 Modelos multiplicativos

Na busca de modelos de regressão múltipla, existe uma liberdade muito grande quanto às variáveis independentes que podem ser usadas. Uma maneira de ampliar ainda a analise é

de usar modelos de regressão múltipla multiplicativos da forma apresentada na equação 2.36 a seguir:

$$Y = k\prod_{i=1}^{n} X_i^{c_i} \qquad \textbf{(2.36)}$$

Onde

Y é a variável dependente (o volume de vendas no nosso caso).

X_i as variáveis independentes ou explicativas.

c_i os coeficientes a serem determinados na construção do modelo.

Esse tipo de modelo pode ser tratado graças à teoria dos modelos de regressão multilineares graças a uma propriedade matemática simples:

$x = \exp(\ln(x))$ para qualquer x diferente de 0.

A partir disso pode-se escrever:

$$\ln(Y) = \ln(k\prod_{i=1}^{n} X_i^{c_i}) = \ln(k) + \ln(\prod_{i=1}^{n} X_i^{c_i}) \qquad \textbf{(2.37)}$$

$$\ln(Y) = \ln(k) + \sum_{i=1}^{n} \ln(X_i^{c_i}) = \ln(k) + \sum_{i=1}^{n} c_i \ln X_i \qquad \textbf{(2.38)}$$

Assim, a equação 2.38 pode ser considerada como uma equação linear caracterizando um modelo multilinear que foi apresentada anteriormente como equação 2.33.

Basta só trabalhar com os logaritmos das variáveis independentes $\ln(X_i)$ para explicar o $\ln(Y)$, que seria a variável dependente do modelo. Para chegar à variável Y que nos interessa, precisa aplicar o exponencial de $\ln(Y)$, uma vez todos os c_i determinados.

Assim, a propriedade matemática $Y = \exp(\ln(Y))$, já introduzida, permite acessar a variável Y da seguinte forma:

$$Y = \exp(\ln(Y)) = \exp\left(\ln(k) + \sum_{i=1}^{n} c_i \ln X_i\right) \quad \textbf{(2.39)}$$

O logaritmo tem a tendência de suavizar a curva do modelo para colar mais à curva real do volume de vendas. Isso se traduze na maioria das vezes em um aumento do valor de R^2 (coeficiente apresentado a seguir) porque o modelo "cola" mais perto da realidade, então tem um poder explicativo melhor.

2.4.4.2.3 Método dos Mínimos Quadrados

Para estimar os coeficientes b_j (j = 1,2,...,k) mais adequados, usa-se o mesmo método utilizado para regressão simples, porém generalizando a k variáveis: o Método dos Mínimos Quadrados.

O objetivo deste método é de minimizar a seguinte expressão:

$$F(\hat{b}_1, \hat{b}_2, \hat{b}_3 ..., \hat{b}_k) = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{b}_1 - \hat{b}_2 X_{2i} - \hat{b}_3 X_{3i} - ... - \hat{b}_k X_{ki})^2 \quad \textbf{(2.40)}$$

com a forma matricial, a equação precedente torna-se:

$$F(\hat{b}_1, \hat{b}_2, \hat{b}_3 ..., \hat{b}_k) = \hat{e}^T \hat{e} \quad \textbf{(2.41)}$$

Após longos cálculos de derivação de F em relação a cada $\hat{b}_j$ e igualando a zero, chega-se à seguinte equação matricial para os estimadores dos coeficientes b_j:

$$\hat{b} = (X^T X)^{-1} X^T Y \quad \textbf{(2.42)}$$

2.4.4.2.4 Medida de ajuste: R^2 e R^2 ajustado

Gujarati (2005), descreve o sentido do coeficiente múltiplo de determinação R^2 como uma medida do quão 'bem' a curva de regressão da amostra se ajusta aos dados. Este coeficiente é calculado com base na análise de variância.

Definem-se três termos de variação em regressão linear:

A Soma dos Quadrados Totais (SQT) que pode ser separada em dois termos: a Soma dos Quadrados Explicada (SQE) e a Soma dos Quadrados dos Resíduos (variação não explicada)

Onde:

$$SQT = SQE + SQR \quad \textbf{(2.43)}$$

$$SQT = \sum_{i=1}^{n} y_i^2 \quad \textbf{(2.44)}$$

$$SQE = \sum_{j=2}^{k} \hat{b}_j \sum_{i=1}^{n} y_i x_{ji} \quad \textbf{(2.45)}$$

$$SQR = \sum_{i=1}^{n} \hat{e}_i^2 \quad \textbf{(2.46)}$$

Com as letras minúsculas, as variações da variável definida em maiúscula.

O coeficiente de determinação é a porcentagem de variações explicadas dentro do total das variações:

$$R^2 = \frac{SQE}{SQT} \quad \textbf{(2.47)}$$

De forma matricial, Pindyck e Rubinfeld, (1998) fornecem a seguinte fórmula (2.48):

$$R^2 = \frac{b^T X^T X b - n\overline{Y}^2}{Y^T Y - n\overline{Y}^2} \qquad \textbf{(2.48)}$$

Observando a equação 2.47, percebe-se que, conforme o número de váriaveis independentes X aumenta, o modelo se torna mais preciso, a proporção de variação explicada aumenta na variação total, então o R^2 aumenta. Assim, Gujarati (2005) expõe que ao comparar dois modelos da mesma variável dependente com números de variáveis independentes diferentes, deve-se tomar cuidado ao comparar os R^2 e levar em conta o número de variáveis independentes consideradas k (com o intercepto). Por esta razão, define-se a seguir o R^2 ajustado, $\overline{R}^2$:

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad \textbf{(2.49)}$$

Segundo Gujarati (2005), uma vez achados os estimadores dos coeficientes b_j, é necessário se assegurar que cada coeficiente é válido, exercendo um Teste-t sobre cada um deles.

2.4.4.2.5 Teste-t

O Teste-t pode ser efetuado caso se recorra à hipótese que o termo de erro e segue uma distribuição normal de média nula e de desvio padrão constante:

$$e \sim N(0, s^2) \qquad \textbf{(2.50)}$$

Uma vez esta hipótese feita, pode-se aplicar o Teste-t sobre cada coeficiente de regressão parcial.

O Teste-t é um teste de hipótese. Sobre cada coeficiente de regressão parciais b_j (j = 1,2,...,k) achados pelo método dos Mínimos Quadrados, testa-se a seguinte hipótese:

$$H_0 : b_j = 0 \; e \; H_1 : b_j \neq 0 \qquad \textbf{(2.51)}$$

A hipótese H_0 verdade seria equivalente a dizer que a variável independente X_j não tem impacto significativo sobre a variável dependente Y com a presença das outras variáveis independentes. O Test-t se baseia sobre a comparação do valor t calculado para o coeficiente com um valor crítico dado pela estatística t para um certo grau de significância.

A equação para o cálculo do t_j do coeficiente b_j é dada a seguir:

$$t_j = \frac{\hat{b}_j}{S(\hat{b}_j)} \qquad \textbf{(2.52)}$$

Onde $\hat{b}_j$ é o estimador de b_j

e $S(\hat{b}_j)$ o desvio padrão de $\hat{b}_j$

Rejeita-se H_0 se $|t_j| > t_{a/2,n-k-1}$ onde a é o grau de significância desejado, n o número de observações e k o número de variáveis independentes. Geralmente o valor usado de a é 0,05 (ou 5%). $t_{a/2,n-k-1}$ é tirado de uma tabela de probabilidade.

2.4.4.2.6 Teste-F

O Teste-F serve para testar a significância global da regressão através um teste de hipótese:

$$H_0 : b_2 = b_3 = ... = b_k = 0$$

H_1: Nem todos os coeficientes de inclinação são simultaneamente zero.

Este Teste-F é bem diferente do Teste-t pois testa o modelo na sua globalidade: um coeficiente b_j pode ser testado pelo Teste-t sem que todos os outros coeficientes de inclinação sejam iguais a zero.

Neste Teste, baseado sobre a análise da variância, vai ser calculado um valor F pela seguinte fórmula 2.53:

$$F = \frac{SQE \big/ (k-1)}{SQR \big/ (n-k)} = \frac{R^2 \big/ (k-1)}{(1-R^2) \big/ (n-k)} \qquad \textbf{(2.53)}$$

Onde (k-1) é grau de liberdade da Soma do Quadrados Explicada

e (n-k) o grau de liberdade da Soma do Quadrados dos Resíduos.

Se $F > F_a(k-1, n-k)$, rejeita-se H_0 e se conclui que, pelo menos, um b_j é diferente de 0. $F_a(k-1, n-k)$ é o valor critico de F em nível de significância a .

2.4.4.2.7 Multicolinearidade

Uma das hipóteses do modelo de regressão multilinear é que não exista relação linear exata entre as variáveis independentes do modelo. Se houver uma relação linear entre variáveis independentes, fala-se que existe colinearidade perfeita.

Existem várias formas de se detectar a multicolinearidade, porém segundo Pindyck e Rubinfeld (1991), nenhuma conquistou de maneira ampla a comunidade científica.

A primeira desta é quando existe um R^2 grande numa equação com valores baixos das razões individuais t. Uma outra maneira é estudar a correlação dois a dois das variáveis independentes. Porém, a existência de uma correlação é suficiente para evidenciar uma, mas não necessária: pode existir colinearidade sem, necessariamente, haver correlação. Neste trabalho, será utilizado o método baseado na análise das correlações entre variáveis.

Será estudada a matriz de correlação de cada modelo para determinar as correlações entre variáveis e assim evidenciar uma colinearidade. Se não há nenhuma correlação, esta análise será considerada suficiente para validar a não colinearidade neste trabalho.

2.4.4.2.8 Estatística de Durbin-Watson

A estatística de Durbin-Watson mede a presença de correlação serial nas variáveis independentes. A correlação serial pode acontecer com uma variável que se correlaciona com uma outra defasada no tempo, ou com ela mesma. Neste último caso a correlação se chama de autocorrelação serial. Com valor da estatística de Durbin-Watson (D-W) perto de 2 não tem presença significada de correlação serial. Com valores entre 2 e 4, existe uma correlação serial negativa. O principal problema acontece quando há evidência de correlação positiva,

quando a estatística está inferior a 1,5. Uma correlação positiva tem as seguintes conseqüências:

- Subestimação dos erros de previsão.
- Superestimação do coeficiente R^2.
- O modelo multilinear vira inválido e pode não ser o melhor método para prever a variável dependente.

Ao longo do nosso trabalho, nós vamos validar unicamente modelos com valor da estatística de Durbin-Watson perto de 2 (entre 1,6 e 3). Esse critério foi baseado no manual de utilização de E-Views 2.0 que nos indica as principais estatísticas calculadas pelo software, assim como as regras de interpretação delas.

2.4.4.3 Método para a resolução do problema

Encontrar um modelo de regressão linear não é uma tarefa simples e seu sucesso não pode ser garantido. Devido ao seu caráter investigativo, esse trabalho pode se tornar muito pesado frente ao gigantesco número de variáveis disponíveis. Não foi possível localizar na revisão bibliográfica algum roteiro que possa ser seguido para se chegar a um modelo de regressão satisfatório, mas sim um conjunto de regras a ser respeitado para assegurar a validade do modelo em desenvolvimento. Isso é uma conseqüência do gigantesco número de variáveis e condições que geram uma grande diversidade de caminhos possíveis para a pessoa em busca de um modelo de regressão linear. Assim será seguido um caminho próprio, onde cada passo depende do último. Serão elaborados vários modelos, cada um tirando os ensinos do precedente para aprimorá-lo.

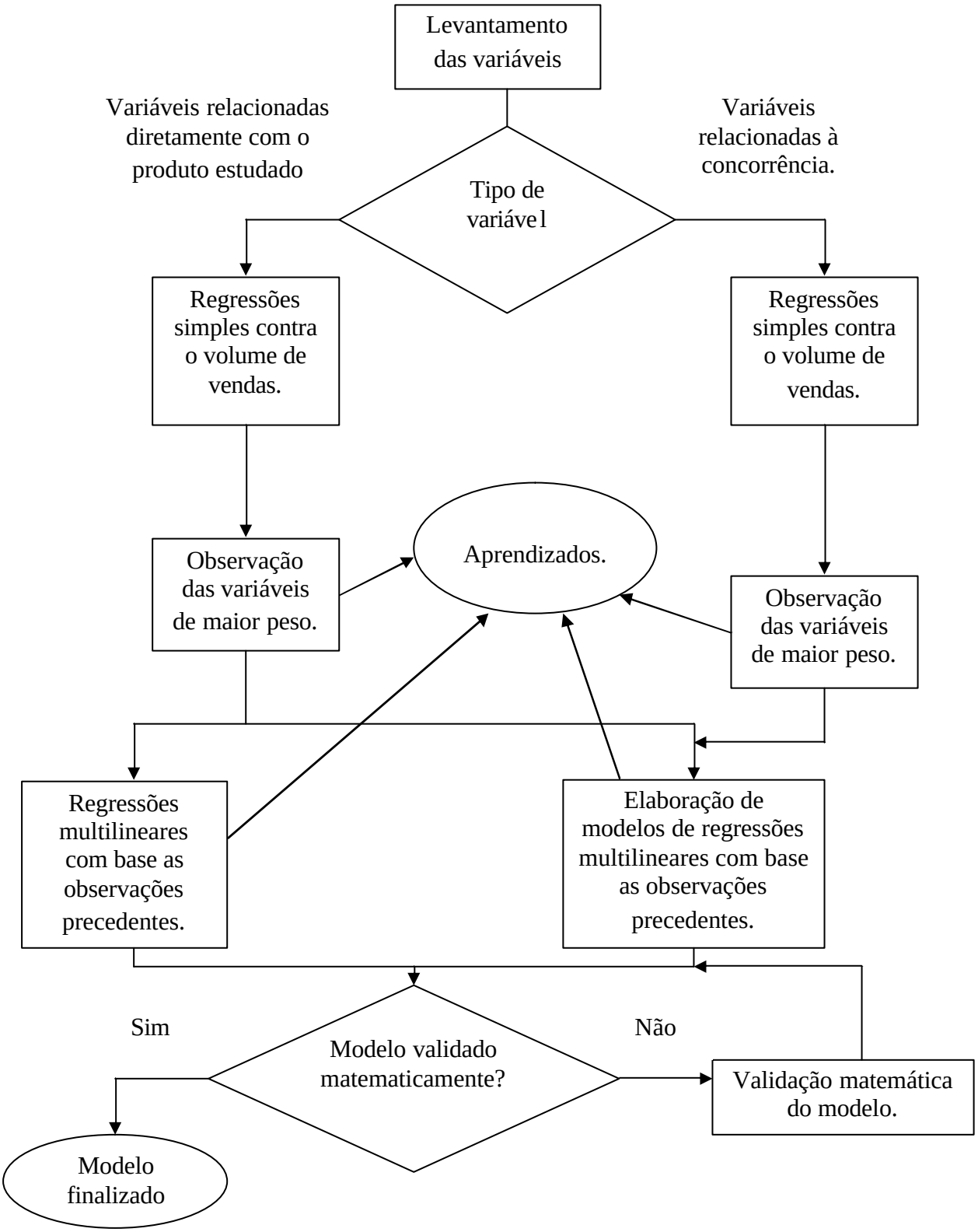
Inicialmente, regredi-se cada variável candidata isoladamente contra a variável dependente, de modo a obter uma primeira idéia de como cada variável relaciona-se individualmente com o volume de vendas. Esse primeiro passo permite adquirir uma sensibilidade em relação à manipulação das variáveis: quais variáveis têm mais peso explicativo? Quais variáveis não têm nenhum peso explicativo em relação ao volume de vendas?... Assim, pode ser elaborada uma primeira triagem das variáveis candidatas para guardar somente as mais pertinentes.

Para isso, vamos proceder em dois tempos. O primeiro, consiste no estudo das variáveis relacionadas ao produto analisado independentemente dos produtos da concorrência e do ambiente competitivo. Após ter discutido com os planejadores e especialistas do mercado de detergente em pó da empresa, foi acordado de regredir as variáveis do produto analisado ignorando os efeitos dos produtos concorrentes. Isso tem como objetivo tirar ensinos sobre o produto da companhia para responder ao tipo de perguntas a seguir: Será que as variações de preço do produto impactam muito o volume de vendas? A distribuição tem um peso significativo no volume de vendas do produto?...

O segundo tempo se interessara às outras variáveis e suas influências sobre o volume de vendas, com o objetivo de guardar as variáveis mais adequadas para a elaboração do nosso modelo.

Por fim, serão elaborados modelos de previsão com base todo esse conhecimento.

O fluxograma 2 a seguir ilustra de maneira geral o caminho para elaborar o modelo de previsão respondendo ao problema do trabalho.



Fluxograma 2 – Método proposto para a realização do trabalho

Frente ao grande número de possibilidade, o raciocínio usado para construir os modelos será uma combinação de intuição e bom senso com os aprendizados das regressões simples.

Os resultados das regressões simples permitem ver quais variáveis têm mais importância para explicar o volume de vendas.

A intuição é usada no momento da escolha de variáveis adicionadas a um modelo para melhorar os seus resultados. As vezes precisa-se privilegiar a abertura do escopo do modelo a um novo tipo de variável do que privilegiar os resultados das variáveis em regressões simples. Quando se fala de abertura de escopo de modelo, pensa-se, por exemplo, em adicionar uma variável de logística a um modelo que conta variáveis de preço e de marketing, mesmo se o desempenho desta variável logística parece menor do que uma outra variável de preço. Com este raciocínio, procura-se buscar novas fontes de informação para melhorar os resultados do modelo.

Por fim, o bom senso é usado se, no modelo estudado, não existe nenhuma variável relacionada ao produto P&G. Acredita-se que um modelo não é completo se ele não leva pelo menos uma variável relacionada ao produto P&G.

No caso de poucas variáveis, podem ser propostos dois métodos para se chegar a um modelo com resultados satisfatórios. O primeiro é de incorporar todas as variáveis disponíveis num modelo e testá-lo estatisticamente para tirar as variáveis menos relevantes. O segundo é de partir da variável de maior R^2 em regressão linear simples e de incorporar a variável que permite o maior aumento do R^2 e assim chegar-se num R^2 superior ou igual a 70%.

Neste trabalho, o desempenho de um modelo é considerado satisfatório se ele é estatisticamente válido, se não existe multicolinearidade e se o seu R^2 é maior do que 70%.

2.4.4.4 Teste dos modelos

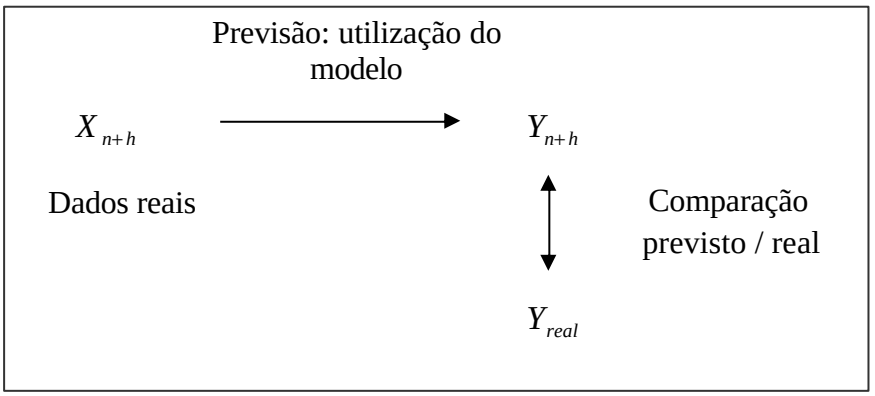
Como visto no esquema 4, uma vez o modelo construído, para usá-lo são necessárias duas etapas de previsão.

A primeira consiste em prever as variáveis independentes X_{n+h}; a segunda, consiste em usar o modelo para prever a variável dependente Y_{n+h}. Neste trabalho, quer-se testar, exclusivamente, o desempenho do modelo sem agregar fontes de erros prevendo os X_{n+h}.

Assim, são utilizados os dados verdadeiros das variáveis independentes para prever a variável dependente de volume de vendas. O modelo é testado sobre dois meses: março e abril.

São comparados os volumes de vendas obtidos através do modelo e os dados reais de março e abril.

O esquema 5, a seguir, esquematiza este processo de teste do modelo multilinear:



Esquema 5 – Teste do modelo

3 Desenvolvimento

3.1 O software E-Views

Como ferramenta para buscar um modelo de regressão multilinear, será usado um software disponibilizado pela empresa P&G: o E-Views 2.0 (Econometric Views 2.0).

Decidiu-se usar esse software, e não o Excel ou o Minitab, por várias razões. A principal é a disponibilidade na empresa do software E-Views 2.0, especializado para apoiar os planejadores a desenvolver modelos de regressão linear. Uma vez que o software tenha sido alimentado com os dados disponíveis de cada variável, tem-se a disposição uma série de módulos para gerar informações necessárias para apoiar as decisões na criação do modelo. Por exemplo, existe uma função que calcula a matriz de colinearidade, uma outra que calcula todos os parâmetros estatísticos, tais como probabilidade do teste t, do teste F etc. Para se chegar nesses mesmos valores com o Excel, demoraria muito para fazer todos os cálculos na mão, sem nenhum valor agregado para o trabalho. O Excel foi usado para a elaboração dos modelos de extrapolação e para toda fase de elaboração e combinação das variáveis dos modelos de regressão, por ser mais fácil de uso na manipulação dos dados. Uma outra razão do uso do E-Views 2.0 foi minha vontade de aprender a usar um software complexo usado pelos planejadores, pois ele oferece possibilidades de cálculos e informações que nem sempre pensamos. E-Views 2.0 permitiu elaborar um trabalho mais completo e profissional. O E-Views 2.0 não possui modulo de escolha das melhores variáveis. Ele calcula todos os parâmetros do modelo que criamos, não escolhendo entre todas as variáveis que estão submetidas a ele, a melhor combinação para ter o modelo mais eficiente possível. Essa característica nos levará a usar os nossos próprios métodos de escolha de variáveis que serão descritos ao longo deste capítulo.

3.2 Descrição das variáveis

Foram levadas em consideração as variáveis relevantes dentro do grupo de informações disponíveis na empresa P&G. A escolha das variáveis, inseridas no escopo deste trabalho, foi feita de acordo com os planejadores e pessoas que têm uma grande experiência no ramo de sabão em pó.

Foram levantadas as variáveis, que serão apresentadas a seguir, tentando sempre recolher o maior histórico possível. Em relação ao histórico, as variáveis de preço foram as que limitaram o estudo ao período de maio 2003 até fevereiro 2006. Armstrong (2001), explica que quanto maior o histórico de dado, maior será a precisão do modelo. Acredita-se que este histórico mensal de quase três anos para cada variável será suficiente para atingir resultados satisfatórios.

A seguir, serão apresentadas as informações que, acredita-se, serão úteis para resolver o problema. Serão apresentadas as fontes dessas informações, assim como o retrabalho que foi feito para se chegar a variáveis exploráveis.

O apêndice A oferece os valores dessas variáveis classificados por produto (P&G e três concorrentes) através dos quadros 17 a 20.

3.2.1 Volume de vendas

A variável dependente que se quer prever é o volume de vendas mensal de um produto de detergente em pó da empresa Procter & Gamble. Esses dados estão fornecidos pela própria empresa P&G em uma determinada unidade e estão multiplicados por um coeficiente para

manter a confidencialidade exigida pela empresa. Essa variável de volume de vendas se chama DVOL.

Para conhecer as características desta variável é feito um histograma da mesma e são calculados os seguintes parâmetros: calculam-se a média, a mediana, o máximo e mínimo desta série de dados. Isto é dado no gráfico 1 a seguir.

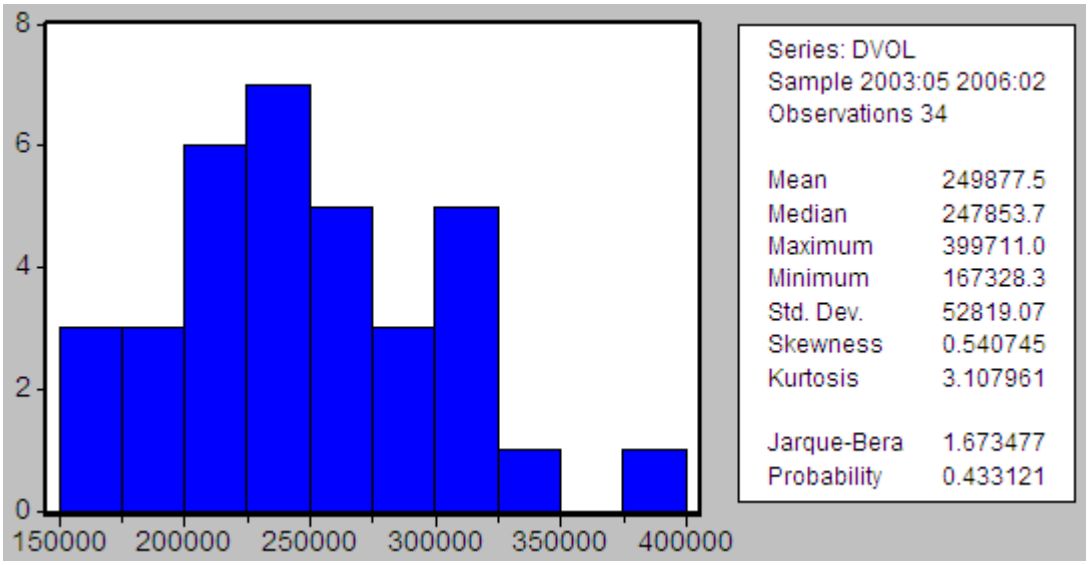


Gráfico 1 – Histograma do volume de venda

3.2.2 Variáveis de Preço

O tipo principal de variável levado em conta é o preço. Essa escolha de levar em conta variáveis de preço vem do fato trivial de que se o preço de um produto aumenta, suas vendas diminuem. Apesar de parecer trivial, esse fenômeno merece ser observado para o produto P&G, assim como o impacto dos preços da concorrência nas vendas do produto P&G para poder tirar aprendizados de quem é o verdadeiro concorrente, quais são as sensibilidades das vendas às essas variáveis etc. Argumentos suplementares para levar em conta variáveis de preço são os resultados das pesquisas do mercado de detergente em pó que mostram a

ocorrência de verdadeiras guerras de preços entre os concorrentes que, às vezes levam-nos a vender a perda. Muitas vezes, o cliente tem como critério de decisão, entre as marcas equivalentes, o preço.

Primeiramente, levantam-se os preços do produto P&G estudado, assim como os preços dos três principais produtos concorrentes. Os dados de preço são levantados pela empresa de pesquisa de mercado AC Nielsen, semanalmente, em várias lojas espalhadas pelo Brasil inteiro. O preço é aquele visto pelo consumidor no momento da compra. Serão usados os dados semanais consolidados do Brasil para o produto P&G e dos três principais concorrentes. Com esses dados semanais, tira-se uma média mensal para se ter uma maneira de elaborar regressões lineares contra a variável de vendas mensais. Para cada um dos produtos estudam-se as suas vendas no tamanho de 1kg. Para este tamanho, existem dois tipos de embalagem: a embalagem em papelão e a embalagem em saco. Uma vez calculadas as médias mensais de cada um desses tipos, são calculadas médias ponderadas pelo volume de vendas de cada formato. Por exemplo, num produto x, vendem-se dois terços de embalagem saco e um terço em papelão. O preço mensal do produto x será a média do preço "papelão" com peso um terço com o preço "saco" com peso dois terços. Para cada um dos quatro produtos será feito esse cálculo.

Assim, chega-se a quatro variáveis de preço mensais por produto. Além disso, são criadas duas variáveis de preço que agrupam vários concorrentes. A primeira é uma variável de preço dos dois principais concorrentes e, a segunda, dos três principais concorrentes. Essas variáveis são calculadas como médias dos preços dos concorrentes ponderados pelo *volume share* de cada um. O *volume share* é um dado fornecido também pela AC Nielsen, que mede qual a participação, em volume (em porcentagem), de cada concorrente em seu mercado, num determinado período. No nosso caso os quatro produtos pertencem ao mesmo mercado porque

competem diretamente, mas a soma dos quatros *volume shares* é inferior a 100% pois existem menores concorrentes neste mercado.

O quadro 1, a seguir, recapitula as variáveis de preço, assim como os seus nomes:

Variável Independente	Símbolo
Preço produto P&G	DPPG
Preço concorrente 1	DC1
Preço concorrente 2	DC2
Preço concorrente 3	DC3
Preço ponderado 3 concorrentes	DP3
Preço ponderado 2 concorrentes (1 e 2)	DP2

Quadro 1 – Variáveis de preço

3.2.3 Índice de preço

A partir das variáveis de preço de cada produto podem-se calcular variáveis combinadas. Criam-se cinco variáveis de índice de preço com o propósito de criar o que acontece na mente do consumidor no momento da compra: a comparação dos preços entre os concorrentes. Além disso, acredita-se que este tipo de variável combinada possui um poder explicativo, no momento da regressão, maior do que uma variável simples. Isso constitui mais uma razão para se criar essas variáveis índices de preço. Essas variáveis são calculadas com o preço do produto P&G, divido pelo preço do(s) produto(s) da concorrência, multiplicado por um fator 100. Assim, serão criadas as cinco variáveis a seguir (quadro 2):

Variável Independente	Símbolo
Índice preço concorrente 1	IC1
Índice preço concorrente 2	IC2
Índice preço concorrente 3	IC3
Índice preço ponderado 3 concorrentes	IP3
Índice preço ponderado 2 concorrentes (1 e 2)	IP2

Quadro 2 – Índices de preço

3.2.4 Distribuição

A distribuição é uma medida de logística que avalia, de maneira grosseira, o estoque do cliente. Existem dois tipos de medidas de distribuição: a distribuição numérica e a distribuição ponderada. A primeira, indica em % o número de lojas que negociaram a marca x durante o ultimo bimestre. Isso é feito para cada marca. Neste trabalho, utiliza-se a distribuição ponderada, que é muito mais representativa da realidade. A distribuição ponderada se calcula com base na distribuição numérica, mas ponderando o resultado de cada loja pelo faturamento da categoria do produto analisado. Por exemplo, se um produto está em distribuição (presente pelo menos uma vez nas últimas oito semanas) numa loja grande, e sem distribuição numa loja pequena, o resultado consolidado para essa duas lojas em distribuição ponderada será muito mais perto da situação “em distribuição” do que o contrário pelo peso maior da loja maior. Esta medida é fornecida também pelo instituto de pesquisa mercadológica AC Nielsen, a cada dois meses.

Usa-se a distribuição ponderada para o produto P&G, assim como para os três concorrentes com os nomes apresentados no quadro 3:

Variável Independente	Símbolo
Distribuição produto P&G	DDIST
Distribuição concorrente 1	DIST1
Distribuição concorrente 2	DIST2
Distribuição concorrente 3	DIST3

Quadro 3 – Variáveis de distribuição

3.2.5 Presença na loja

A presença na loja é também uma medida de logística, mas muito mais voltada ao consumidor. Se o produto estiver em distribuição na loja visitada pode ser medida a presença na loja do produto, caso contrário, a presença na loja não será medida.

Para se medir a presença na loja, há necessidade de se conhecer os hábitos do consumidor. Basta ir na prateleira e ver se o produto está presente. Assim, ponderando essas presenças binárias nas lojas pelo faturamento da categoria do produto analisado, chega-se a uma percentagem que será usada para elaborar os nossos modelos.

As quatro variáveis de presença na loja usadas são as seguintes (quadro 4):

Variável Independente	Símbolo
Presença na loja produto P&G	DPRE
Presença na loja concorrente 1	DPRE1
Presença na loja concorrente 2	DPRE2
Presença na loja concorrente 3	DPRE3

Quadro 4 – Variáveis de presença na loja

3.2.6 Ponto de Venda (PDV)

Essa variável faz a conta, mensalmente, de todo e qualquer material promocional ou publicitário colocado temporariamente nas lojas e que se refere a produtos específicos e não genericamente. Segundo a definição da AC Nielsen, estão incluídos cartazes (fixados nos diversos locais do estabelecimento), cantoneiras, faixas de gôndolas ou especiais, forrações de gôndolas ou de material exposto em locais em destaque, móbiles etc. Estes dados são medidos a cada dois meses pela AC Nielsen.

As três variáveis usadas estão citadas a seguir, no quadro 5:

Variável Independente	Símbolo
PDV produto P&G	DPDV
PDV concorrente 1	DPDV1
PDV concorrente 2	DPDV2

Quadro 5 – Variáveis de Ponto De Venda (PDV)

O concorrente 3 tem essa variável de ponto de venda zerada quase todos os meses. Ela não será levada em conta para a elaboração do modelo.

3.2.7 Pontos Extras de Armazenamento (PEA)

Essa variável mede, mensalmente, o número de locais, diferentes aos comuns de armazenamento, em que são exibidos os produtos. O objetivo desses pontos extras é destacar para o público-alvo os produtos na loja. As variáveis no quadro 6 serão usadas neste trabalho:

Variável Independente	Símbolo
Pontos extras produto P&G	DEP
Pontos extras concorrente 1	DEP1
Pontos extras concorrente 2	DEP2
Pontos extras concorrente 3	DEP3

Quadro 6 – Variáveis de pontos extras

3.2.8 Logaritmo das variáveis

Todas essas variáveis têm um espelho em logaritmo neperiano. Para cada variável será aplicado o logaritmo neperiano ln para abrir o escopo dos nossos modelos aos modelos multiplicativos, que foram apresentados na revisão bibliográfica.

O quadro 21, em apêndice B, lista estas variáveis em logaritmo neperiano.

3.3 Resultados dos métodos de regressão linear

3.3.1 Modelos lineares com base nas variáveis relativas ao produto P&G

Num primeiro momento, regrediam-se cada variáveis do produto P&G contra o volume de vendas para se ter um pouco de sensibilidade com as variáveis manipuladas.

3.3.1.1 Modelos lineares simples – Testes das variáveis

Antes de apresentar os resultados propriamente ditos, será necessário estabelecer quais os critérios de sucesso de um modelo de regressão linear foram levados em conta. Já foram apresentados, na revisão bibliográfica, os testes estatísticos necessários à validação de um modelo, assim como as medidas que nos ajudam a saber se o modelo está satisfatório. Os principais parâmetros verificados em cada modelo foram o coeficiente de determinação R^2, o R^2 ajustado, valor P do teste t, valor P do teste F e o erro padrão de regressão (em %). Como sabemos, para um modelo de regressão linear simples, ou seja, com somente uma variável independente, o teste t é idêntico ao teste F.

Logo a seguir (Figura 1), pode-se ver o relatório de resultados do software E-Views após a regressão da variável de preço do produto P&G estudado contra o volume de vendas. Cercados de vermelho, aparecem os parâmetros descritos anteriormente, que servirão para avaliar a pertinência das variáveis para prever o volume de vendas.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DPPG	141469.7	38832.13	3.643109	0.0009
C	-462118.7	195589.4	-2.362698	0.0244

R-squared	0.293165	Mean dependent var	249877.5
Adjusted R-squared	0.271077	S.D. dependent var	52819.07
S.E. of regression	45095.35	Akaike info criterion	21.49009
Sum squared resid	6.51E+10	Schwarz criterion	21.57988
Log likelihood	-411.5755	F-statistic	13.27224
Durbin-Watson stat	1.662827	Prob(F-statistic)	0.000943

Figura 1 – Parâmetros da regressão simples do preço do produto P&G

Verifica-se que o teste F é equivalente ao teste t no caso de um modelo de regressão simples, pois a probabilidade F está igual à probabilidade t.

De maneira mais visual, pode-se observar, no gráfico 2, o quanto o resultado do modelo linear simples (curva *Fitted*) "cola" a curva de volume de vendas (*Actual*). A curva residual permite seguir a evolução do erro entre *Actual* e *Fitted*.

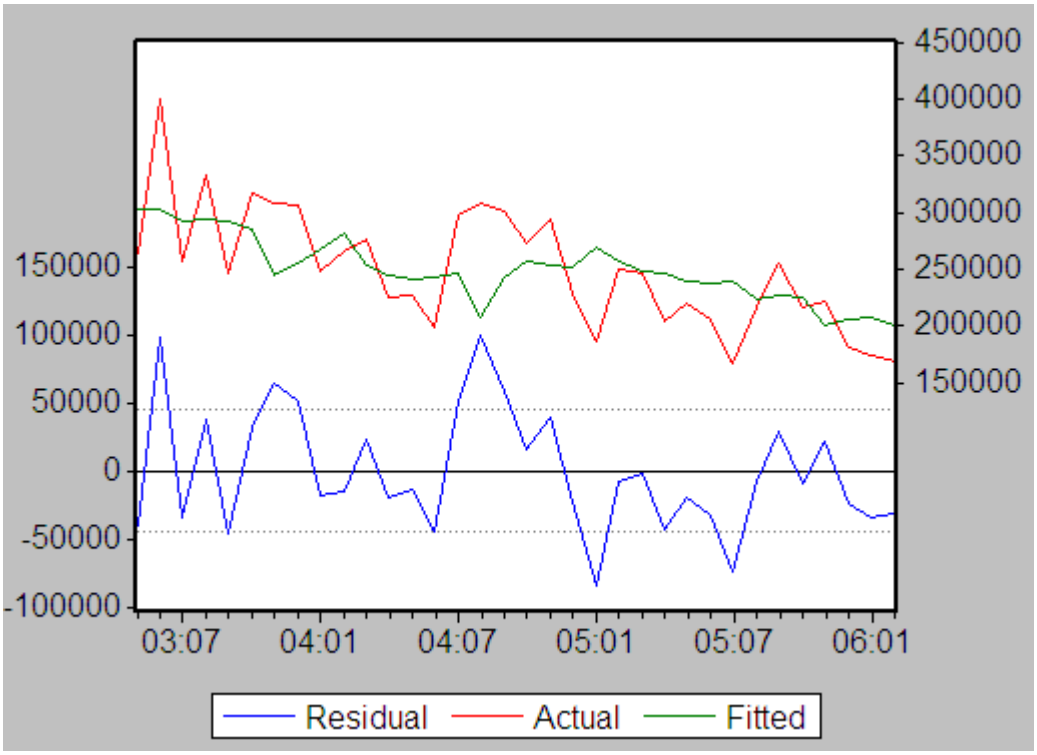


Gráfico 2 – Gráfico da regressão simples do preço do produto P&G

No quadro 7 estão sumarizados os resultados das regressões lineares simples das variáveis relacionadas ao produto P&G, contra o volume de vendas desse produto.

Variável Independente	Símbolo	R^2	R^2 Ajustado	Valor P do teste t	Erro Padrão
Preço produto P&G	DPPG	0.293	0.271	0.0009	18.0%
Distribuição produto P&G	DDIST	0.402	0.383	0.0001	16.6%
Presença na Loja produto P&G	DPRE	0.031	0.001	0.3166	21.1%
PDV produto P&G	DPDV	0.113	0.086	0.0514	20.2%
Pontos extras produto P&G	DEP	0.043	0.013	0.2372	21.0%

Quadro 7 – Resultados das regressões lineares simples (produto P&G)

À luz desses resultados, pode-se perceber que as duas variáveis com mais poder explicativo para a previsão do volume de vendas são o preço do produto e a distribuição desse produto na loja. Um primeiro aprendizado é que a distribuição do produto parece ter mais peso na explicação do volume de vendas do que o preço do produto. Isso poderia ser uma surpresa conhecendo o ambiente altamente competitivo do mercado de sabão em pó, onde, segundo as pesquisas ao consumidor, o posicionamento do preço tem um peso fundamental nas vendas do produto.

3.3.1.2 Modelos multilineares e resultados

Uma vez cada variável do produto P&G testada separadamente contra a variável volume de vendas, serão elaborados modelos de regressão multilinear.

Para a seleção das variáveis que vão entrarão no modelo, a situação ideal seria ter um módulo do software E-Views que selecionasse o grupo de variáveis mais adequado para prever o volume de vendas. Infelizmente, a nossa versão do software não possui tal algoritmo de seleção do melhor grupo de variável.

Como o número de variáveis relativas ao produto P&G não é tão grande serão usados dois métodos intuitivos de escolha das variáveis. Vale frisar, neste ponto do trabalho, que serão integradas nos modelos as variáveis defasadas de até dois períodos para trás. Uma variável X defasada de i apareça como X(-i) nas nossas anotações.

3.3.1.2.1 Modelo 1

O primeiro método, desenvolvido a seguir, é baseado no estudo anterior das regressões simples. Primeiramente será integrada a variável de maior R^2. Num segundo, será integrada a variável que aumenta mais o R^2 e assim por diante até chegar ao maior R^2. Por fim, serão estudados os testes estatísticos (teste t, teste F, Durbin Watson...) para tirar as variáveis inadequadas e validar o modelo. A cada passo, uma vez a melhor variável adicionada ao modelo, junto com as mesmas defasadas de um e dois meses, serão eliminadas, entre essas três mesmas variáveis defasadas, a(s) que estiver(em) muito inapropriadas: com um valor de t muito baixo em valor absoluto.

O resultado obtido está resumido na Figura 2 abaixo. O modelo, que será chamado de modelo 1, está composto de quatro variáveis independentes após aplicado o teste t. Segundo a tabela da estatística de Student, os valores dos coeficientes *t-Statistic* devem estar superiores a 1,8 para validar com uma precisão de mais de 95% o coeficiente da variável correspondente. O erro padrão desse modelo é de 12,6%.

O teste de Durbin-Watson para testar a correlação serial mostra um valor um pouco superior a 2, o que valida a ausência deste fenômeno. O teste F evidencia que pelo menos um coeficiente não está nulo. Neste modelo, o valor de R^2 atingindo é 62,7%, que é um valor razoável para um modelo de somente quatro variáveis referentes unicamente ao produto P&G.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DDIST[-2]	21156.84	3725.704	5.678616	0.0000
DPRE	11587.66	3920.848	2.955398	0.0064
DPDV[-1]	6907.562	3063.393	2.254873	0.0325
DEP[-2]	-8781.669	3062.971	-2.867042	0.0079
C	-2656092.	589870.2	-4.502841	0.0001

R-squared	0.626864	Mean dependent var	244750.8
Adjusted R-squared	0.571584	S.D. dependent var	47031.88
S.E. of regression	30784.00	Akaike info criterion	20.81210
Sum squared resid	2.56E+10	Schwarz criterion	21.04112
Log likelihood	-373.3997	F-statistic	11.33991
Durbin-Watson stat	2.118937	Prob(F-statistic)	0.000016

Figura 2 – Parâmetros do Modelo 1

O gráfico 3 a seguir mostra, de maneira mais visual, o comportamento dos valores do modelo em relação aos valores reais.

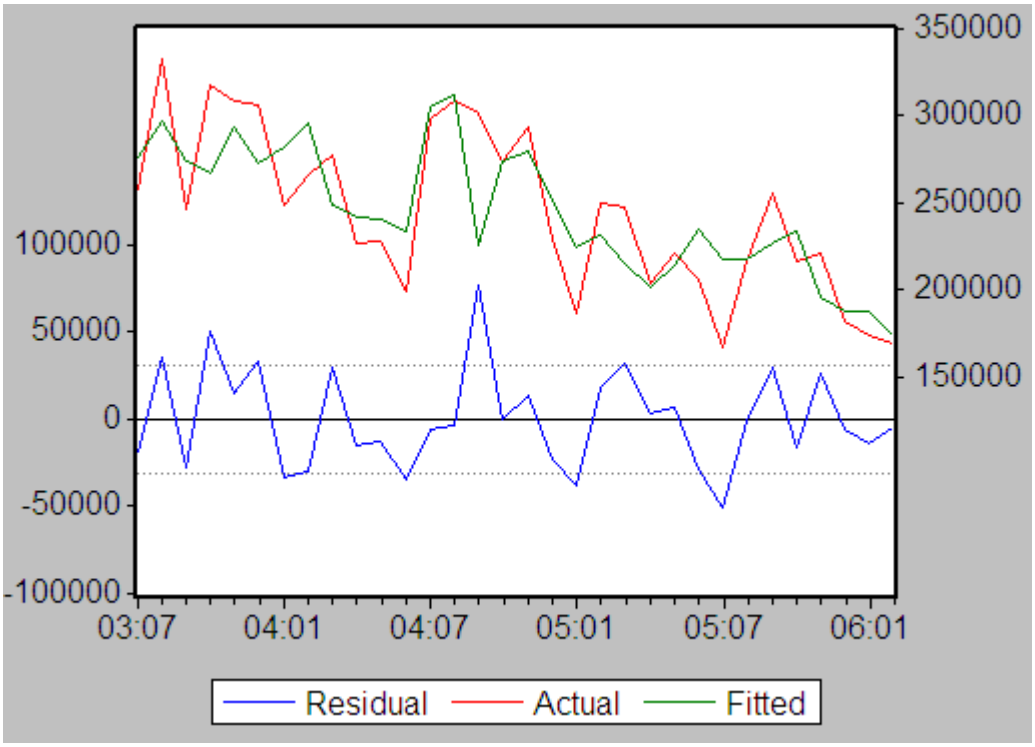


Gráfico 3 – Gráfico do Modelo 1

Um ponto importante a destacar neste ponto do trabalho é o impacto do uso de uma variável defasada. Neste modelo, usam-se duas variáveis defasadas de dois meses: DDIST(-2) e DEP(-2). Isso tem conseqüências sobre o período coberto pelo modelo. Neste caso, foi reduzido o escopo temporal do modelo de dois meses. Os dois primeiros meses das variáveis não têm valores dois períodos atrás, o que impossibilita o uso desses dois primeiros períodos. Temos que saber que essa "supressão" dos dois primeiros valores tem um impacto no modelo como um todo: sobre o R^2, os coeficientes, etc.

3.3.1.2.2 Modelo 2

O segundo método, bem intuitivo também, oferece um caminho diferente. Ele vem do fato de que uma variável pode não ter um R^2 muito alto em regressão simples, mas junto com outras variáveis pode ter um poder explicativo grande, assim como um coeficiente t bem alto. Essa característica foi bem comprovada com a variável de presença na loja DPRE no modelo precedente. Em regressão linear simples, essa variável tem o menor R^2 das cinco variáveis independentes, mas junto com outras variáveis ela tem um poder explicativo alto e um valor do coeficiente t alto. Serão integradas, no começo, todas as variáveis ao modelo e aplicados os testes estatísticos (principalmente o teste t) para tirar as variáveis inadequadas.

Os resultados do modelo estão apresentados na figura 3 a seguir. Com o método apresentado precedentemente, chega-se a um modelo, que será chamado modelo 2, composto de sete variáveis independente. O coeficiente R^2 chegou num valor mais alto do que o modelo precedente: 72,0%. Da mesma maneira os valores dos coeficientes têm que estar superiores a 1,8 para-se ter certeza de que o coeficiente da variável correspondente está diferente de 0, com 95% de certeza. O erro padrão neste caso é reduzido por 11,6%.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DPPG(-2)	80543.26	37337.44	2.157172	0.0412
DDIST(-2)	19248.59	4204.063	4.578569	0.0001
DPRE	11250.10	3749.927	3.000086	0.0062
DPDV(-1)	6311.136	3356.717	1.880151	0.0723
DPDV	7926.995	3829.926	2.069751	0.0494
DEP(-2)	-10893.17	2985.406	-3.648807	0.0013
DEP	-6222.308	3068.559	-2.027762	0.0538
C	-2816253.	571008.8	-4.932065	0.0000

R-squared	0.720096	Mean dependent var	244750.8
Adjusted R-squared	0.638458	S.D. dependent var	47031.88
S.E. of regression	28279.50	Akaike info criterion	20.71210
Sum squared resid	1.92E+10	Schwarz criterion	21.07854
Log likelihood	-368.7997	F-statistic	8.820545
Durbin-Watson stat	2.378044	Prob(F-statistic)	0.000023

Figura 3 – Parâmetros do Modelo 2

De maneira mais visual, podem-se comparar os resultados do modelo com a realidade do volume de vendas no gráfico 4 a seguir:

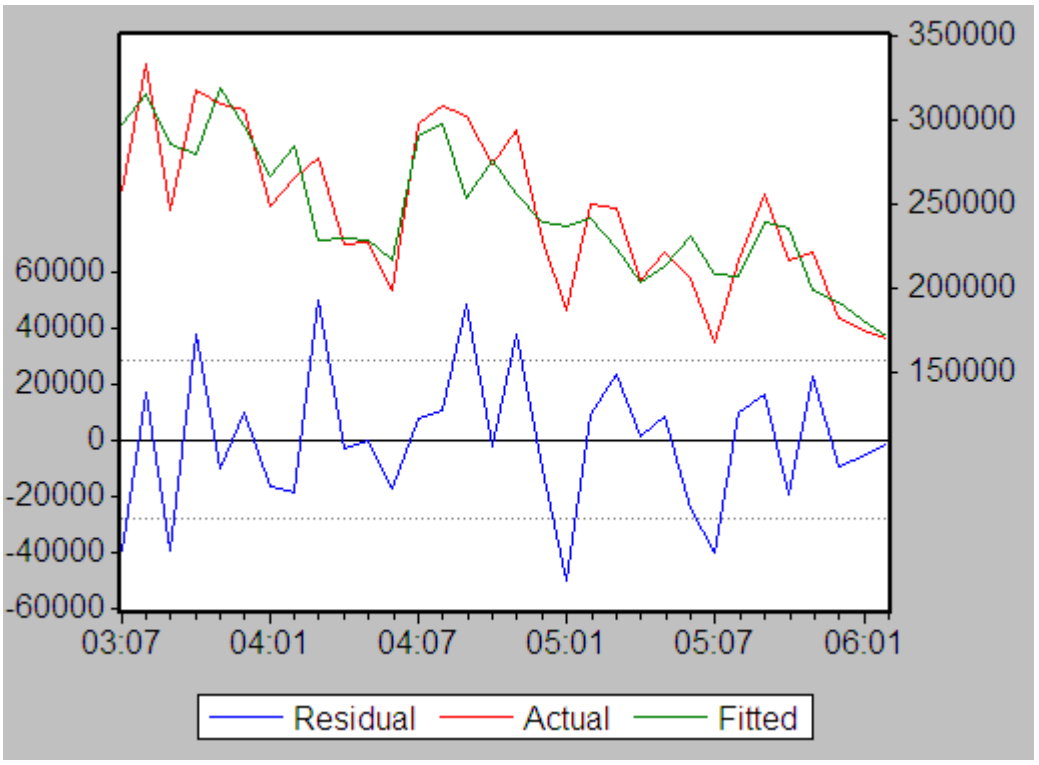
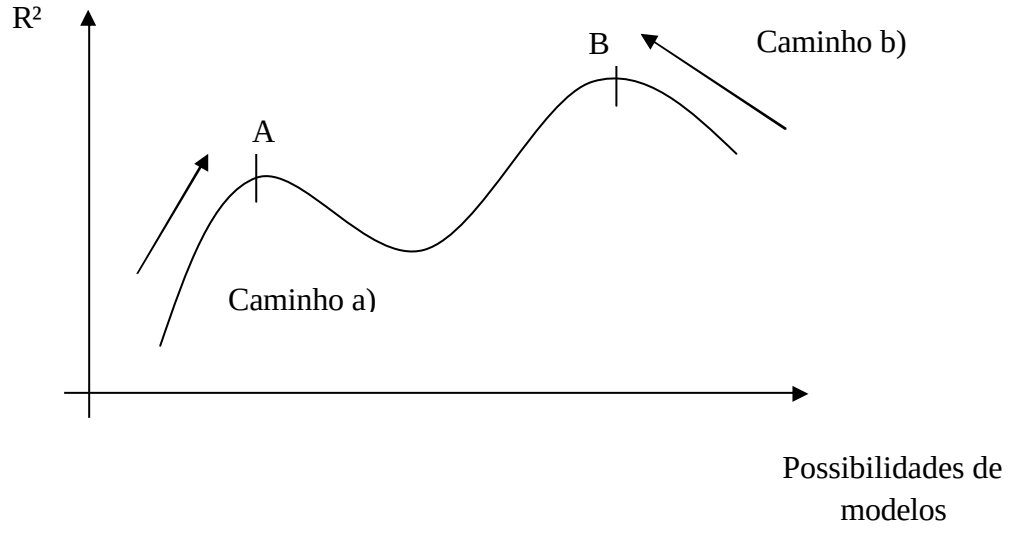


Gráfico 4 - Gráfico do Modelo 2

Neste momento, surge uma pergunta bem legítima na mente do leitor: por que dois modelos tão diferentes com base os mesmos dados? Os dois raciocínios não deveriam levar ao mesmo modelo otimizado com o mesmo R^2?

Já foram explicados os dois métodos usados para se chegar nestes modelos. O primeiro, seria qualificado de "agregado" porque ele agrega variável após variável até verificar os testes estatísticos. O segundo, tem um método diferente que poderia ser qualificado de "reverso": agregam-se todas as variáveis e tiram-se, uma por uma, as variáveis não apropriadas por causa dos testes estatísticos. Esses, são somente dois caminhos numa grande quantidade de caminhos. Existem, no total, $2^{15}-1$ ou 32 767 possibilidades de modelos diferentes (5 variáveis com 3 possibilidades de defasagem). Para explicar qualitativamente a diferença entre os dois modelos obtidos com dois métodos, que, intuitivamente, levariam na melhor solução, podemos tentar nos ajudar com o esquema 6 seguinte:



Esquema 6 – Comparação ilustrativa dos dois modelos

Pode-se ver que no caminho: a) temos um raciocínio que nos leva a aumentar o valor de R^2 até o máximo A; se continuarmos, vamos diminuir o R^2 ; então paramos em A. Com um outro caminho, o b), elevamos o R^2 até ele começar a diminuir de novo; então, paramos em B. A diferença está em que A é um máximo local quando B é o máximo global. Na verdade, nunca saberemos se atingimos o máximo global ou um máximo local. A única certeza que nós temos nos nossos dois modelos é que o primeiro atingiu um máximo local de R^2. Temos que observar que esse gráfico simples em duas dimensões é válido porque temos dois caminhos, mas, se escolhemos n caminhos teria que aumentar o número de dimensão do desenho... Essa explicação qualitativa frisa bem a complexidade da elaboração de um modelo de regressão multilinear e a vantagem de se ter um software que faz os cálculos de todas as possibilidades (que com cinco variáveis já estão de 32 767) para achar o melhor modelo. Como já explicado anteriormente, o E-Views 2.0 não possui tal algoritmo.

3.3.1.2.3 Análise de multicolinearidade

Para validar os modelos, precisa-se analisar a multicolinearidade dentro das variáveis selecionadas. Para isso, serão usadas matrizes de correlação, como já explicado na revisão bibliográfica.

A matriz de correlação do modelo 1 está apresentada a seguir (no quadro 8) e pode-se reparar que o maior valor em módulo dela é uma correlação entre a variável dependente de volume e a variável independente de distribuição. Isso valida o fato de não ter multicolinearidade entre as variáveis do modelo.

	DDIST(-2)	DEP(-2)	DPDV(-1)	DPRE	DVOL
DDIST(-2)	1.000000	-0.367820	-0.584797	-0.424912	0.671840
DEP(-2)	-0.367820	1.000000	0.487953	0.440070	-0.396135
DPDV(-1)	-0.584797	0.487953	1.000000	0.148832	-0.310892
DPRE	-0.424912	0.440070	0.148832	1.000000	-0.094009
DVOL	0.671840	-0.396135	-0.310892	-0.094009	1.000000

Quadro 8 – Matriz de correlação do Modelo 1

O mesmo cálculo de matriz de correlação foi feito para as variáveis do modelo 2. O resultado está apresentado no quadro 9 a seguir.

	DDIST(-2)	DEP	DEP(-2)	DPDV	DPDV(-1)	DPPG(-2)	DPRE	DVOL
DDIST(-2)	1.000000	-0.256517	-0.367820	-0.608649	-0.584797	0.593828	-0.424912	0.671840
DEP	-0.256517	1.000000	0.005372	0.606134	0.358302	-0.057038	0.114845	-0.119294
DEP(-2)	-0.367820	0.005372	1.000000	0.295885	0.487953	-0.082357	0.440070	-0.396135
DPDV	-0.608649	0.606134	0.295885	1.000000	0.708163	-0.454523	0.155972	-0.278605
DPDV(-1)	-0.584797	0.358302	0.487953	0.708163	1.000000	-0.453152	0.148832	-0.310892
DPPG(-2)	0.593828	-0.057038	-0.082357	-0.454523	-0.453152	1.000000	-0.006193	0.537240
DPRE	-0.424912	0.114845	0.440070	0.155972	0.148832	-0.006193	1.000000	-0.094009
DVOL	0.671840	-0.119294	-0.396135	-0.278605	-0.310892	0.537240	-0.094009	1.000000

Quadro 9 – Matriz de correlação do Modelo 2

Nesta matriz, o maior valor em módulo existe entre duas variáveis independentes DPDV e DPDV(-1). Isso evidencia a presença de multicolinearidade neste modelo. Segundo Armstrong (1985), nesta situação existem duas possibilidades:

- Suprimir uma das duas variáveis do modelo para tirar essa multicolinearidade do modelo.

- Criar uma nova variável combinando as duas variáveis em causa.

Essa segunda solução é privilegiada por Armstrong (1985), porque ela guarda a maior parte da informação útil à boa precisão do modelo.

Antes de criar uma nova variável, precisa-se parar um pouco para entender por que houve colinearidade entre DPDV e DPDV(-1). A primeira idéia que vem em mente é que

trata-se da mesma variável defasada de um mês. Mas, olhando de mais perto o modelo, pode-se ver que entre DEP e DEP(-2) não existe quase nenhuma colinearidade. Na verdade, a causa da colinearidade entre DPDV e DPDV(-1) se situa no método de medição desses dados pela AC Nielsen. Foi explicado na definição das variáveis que esses dados eram medidos bimestralmente. Assim, a cada 2 meses o valor de DPDV é igual ao dados do mês anterior. Desta maneira, claramente existe colinearidade entre DPDV e a mesma variável defasada de um mês.

3.3.1.2.4 Modelo 3

Para remediar esse problema de multicolinearidade no modelo 2, será criada uma variável nova a partir de DPDV e DPDV(-1). Ela é a soma de DPDV com DPDV(-1). Na figura 4 a seguir, chamado de modelo 3, ela aparece como DPDV.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DPPG(-2)	80219.50	36619.77	2.190606	0.0380
DDIST(-2)	19173.86	4116.440	4.657874	0.0001
DPRE	11300.03	3675.273	3.074610	0.0050
DPDV	7042.269	1948.467	3.614262	0.0013
DEP(-2)	-11046.50	2876.104	-3.840786	0.0007
DEP	-5956.529	2852.166	-2.088423	0.0471
C	-2813084.	560203.9	-5.021536	0.0000

R-squared	0.719245	Mean dependent var	244750.8
Adjusted R-squared	0.651864	S.D. dependent var	47031.88
S.E. of regression	27750.23	Akaike info criterion	20.65264
Sum squared resid	1.93E+10	Schwarz criterion	20.97327
Log likelihood	-368.8483	F-statistic	10.67428
Durbin-Watson stat	2.385903	Prob(F-statistic)	0.000007

Figura 4 – Parâmetros do Modelo 3

Pode-se ver que o R^2 quase não diminui, significando, neste modelo que foi guardado o mesmo poder explicativo que no modelo 2. O erro padrão diminui um pouco para atingir 11,3%.

O gráfico 5 a seguir, permite acompanhar a evolução do erro, diferença entre a curva real e a curva do modelo.

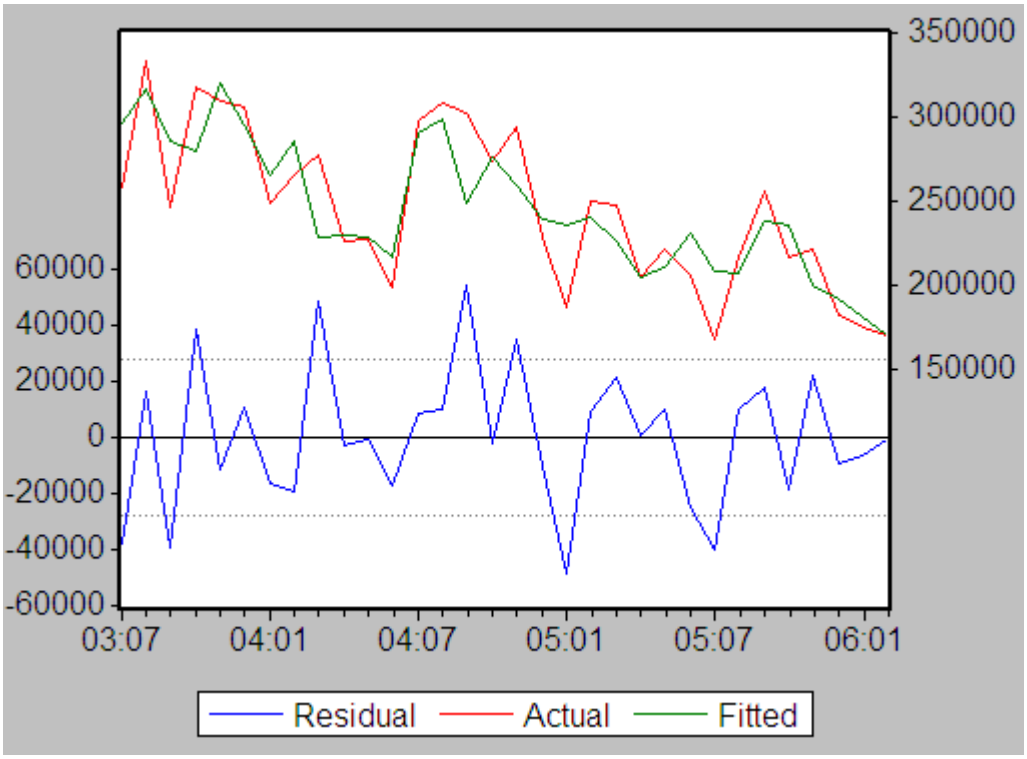


Gráfico 5 – Gráfico do Modelo 3

Apesar de não apresentar grande necessidade, será recalculada a matriz de correlação (ou pelo menos os novos coeficientes relacionados com a nova variável DPDV) para se ter certeza de que não sobra multicolinearidade.

	DDIST(-2)	DEP	DEP(-2)	DPDV	DPPG(-2)	DPRE	DVOL
DDIST(-2)	1.000000	-0.256517	-0.367820	-0.645561	0.593828	-0.424912	0.671840
DEP	-0.256517	1.000000	0.005372	0.520507	-0.057038	0.114845	-0.119294
DEP(-2)	-0.367820	0.005372	1.000000	0.425065	-0.082357	0.440070	-0.396135
DPDV	-0.645561	0.520507	0.425065	1.000000	-0.491007	0.164870	-0.319098
DPPG(-2)	0.593828	-0.057038	-0.082357	-0.491067	1.000000	-0.006193	0.537240
DPRE	-0.424912	0.114845	0.440070	0.164870	-0.006193	1.000000	-0.094009
DVOL	0.671840	-0.119294	-0.396135	-0.319098	0.537240	-0.094009	1.000000

Quadro 10 – Matriz de correlação do Modelo 3

Essa matriz (quadro 10) tem o seu maior coeficiente entre a variável dependente de volume e uma variável independente DDIST(-2). Isso significa que não existe evidência de multicolinearidade nesse modelo.

3.3.2 Modelos lineares com base todas as variáveis

3.3.2.1 Modelos lineares simples das variáveis externas ao produto P&G

Procede-se da mesma maneira que para as variáveis relativas ao produto P&G. Num primeiro momento, regredi-se cada variável independente contra o volume de vendas, para sentir um pouco quais são as variáveis de maior peso, mesmo se já foi visto nos modelos anteriores que uma variável de peso fraco, num modelo de regressão simples, pode ter um peso bem maior num modelo multilinear.

Os resultados para cada variáveis são apresentados no quadro 11 a seguir:

Variável Independente	Símbolo	R^2	R^2 Ajustado	Valor P do teste t	Erro Padrão
Preço Concorrente 1	DC1	0.132	0.105	0.0344	20.0%
Preço Concorrente 2	DC2	0.224	0.200	0.0047	18.9%
Preço Concorrente 3	DC3	0.315	0.294	0.0005	17.8%
Preço ponderado 3 concorrentes	DP3	0.298	0.276	0.0008	18.0%
Preço ponderado 2 concorrentes	DP2	0.170	0.144	0.0154	19.6%
Índice preço concorrente 1	IC1	0.061	0.031	0.1591	20.8%
Índice preço concorrente 2	IC2	0.004	(0.027)	0.7271	21.4%
Índice preço concorrente 3	IC3	0.000	(0.041)	0.9171	18.4%
Índice preço ponderado 3 concorrentes	IP3	0.035	0.005	0.2903	21.1%
Índice preço ponderado 2 concorrentes	IP2	0.031	0.001	0.3180	21.1%
Distribuição concorrente 1	DIST1	0.000	(0.031)	0.9991	21.5%
Presença na loja concorrente 1	DPRE1	0.129	0.101	0.0372	20.0%
PDV concorrente 1	DPDV1	0.046	0.017	0.2209	21.0%
Pontos Extras concorrente 1	DEP1	0.008	(0.023)	0.6214	21.4%
Distribuição concorrente 2	DIST2	0.243	0.220	0.0030	18.7%
Presença na loja concorrente 2	DPRE2	0.135	0.108	0.0328	20.0%
PDV concorrente 2	DPDV2	0.047	0.017	0.2170	21.0%
Pontos Extras concorrente 2	DEP2	0.019	(0.011)	0.4346	21.3%
Distribuição concorrente 3	DIST3	0.425	0.407	0.0000	16.3%
Presença na loja concorrente 3	DPRE3	0.077	0.048	0.1132	20.6%
Pontos Extras concorrente 3	DEP3	0.049	0.019	0.2082	20.9%

Quadro 11 – Resultados das regressões lineares simples (Produto da concorrência)

Têm-se variáveis de peso forte de vários tipos: A distribuição do concorrente 3, DIST3, é uma variável logística e tem o maior peso explicativo da variável dependente de volume de vendas, com um R^2 de 0,425. Variáveis de natureza mais econômicas, tais como DC3 e DP3, têm também, pesos importantes com coeficientes R^2 respectivos de 0,315 e 0,298.

3.3.2.2 Modelos multilineares convencionais

Armstrong (2002), frisa um ponto importante da teoria econométrica: ao elaborar um modelo de regressão multilinear deve-se pensar em obter um modelo simples, com um número de variável não muito elevado.

3.3.2.2.1 Modelo 4

Assim, pode-se apresentar, a seguir (figura 5) um modelo, que se chamará de modelo 4, constituído de duas variáveis que apresentam resultados de qualidade média, em relação à acuracidade da previsão mas que, pelo menos, têm o mérito de ser muito simples a implementar. Precisa-se, somente, de duas variáveis independentes: o preço do concorrente 1 defasado de um mês e a distribuição do concorrente 3. Foram utilizados os resultados do estudo anterior de regressão linear simples de cada variável para determinar as variáveis mais adequadas para elaborar esse modelo. DIST3 e DC1 são as duas variáveis de maior coeficiente R^2 na tabela precedente.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DC1(-1)	73658.68	5842.325	12.60777	0.0000
DIST3	-3259.783	612.0600	-5.325920	0.0000

R-squared	0.537762	Mean dependent var	249446.6
Adjusted R-squared	0.522851	S.D. dependent var	53577.31
S.E. of regression	37009.06	Akaike info criterion	21.09653
Sum squared resid	4.25E+10	Schwarz criterion	21.18723
Log likelihood	-392.9177	F-statistic	36.06497
Durbin-Watson stat	1.658089	Prob(F-statistic)	0.000001

Figura 5 – Parâmetros do Modelo 4

Um ponto interessante é o alto valor dos coeficientes t das duas variáveis dependentes, significando que estas duas variáveis são muito relevantes para este modelo. Da mesma maneira, a probabilidade de que os dois coeficientes sejam simultaneamente zerados é quase nula, como o indica a probabilidade F.

Neste modelo, não se usou interceptou C porque ele diminuía o coeficiente R^2. O coeficiente R^2 se elevou a 53,7% e o erro padrão do modelo é de 14,8%. Com duas únicas variáveis, o resultado é muito interessante, porém, neste estudo, ainda insuficiente.

A seguir (gráfico 6), apresenta-se o desempenho do modelo de maneira visual. Pode-se observar que a curva calculada pelo modelo não varia muito e somente segue a tendência do volume de vendas. Assim, há necessidade de se achar um modelo com poder explicativo maior para tentar explicar as variações mensais do volume de vendas.

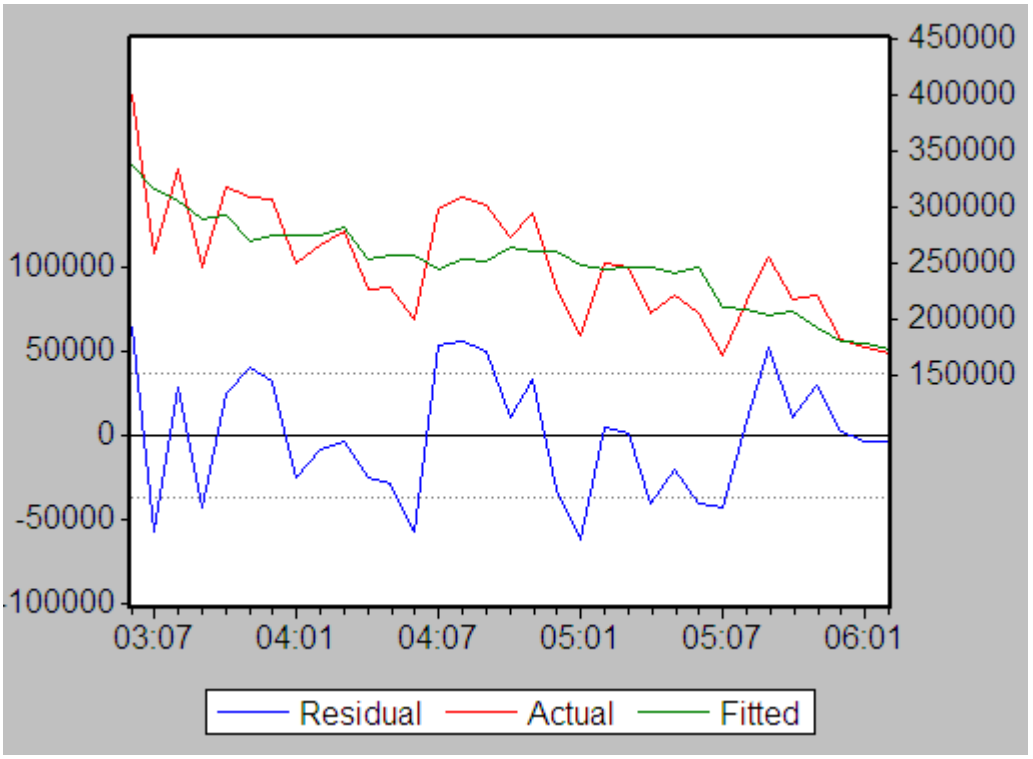


Gráfico 6 – Gráfico do Modelo 4

3.3.2.2.2 Modelo 5

Sempre guardando em mente o conselho de Armstrong (2002) sobre a simplicidade dos modelos que deve ser respeitada, apresenta-se o modelo 5, a seguir, que se compõe de três variáveis. Da mesma maneira que no modelo 4, a escolha das variáveis está feita de acordo com o desempenho das variáveis independentes em regressão simples, cujo resultados são resumidos na tabela precedente. Assim, pode-se observar que a variável DIST3 com o seu R^2 de 0,425 esta de novo presente no modelo. A principal diferença entre o modelo 5 e o 4 é a abertura do seu escopo a novas informações. O modelo 4 não leva em conta informações sobre o produto P&G, nem sobre o produto 2. Neste modelo 5, o produto P&G aparece através da variável IC1 que é o índice entre os preços do produto P&G e do concorrente 1. O produto 2 aparece, também, no modelo através da sua variável de distribuição defasada de um mês.

Os resultados do modelo 5 estão resumidos na figura 6 a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
IC1	-8626.890	2684.840	-3.213186	0.0031
DIST3	-4818.783	754.2951	-6.388459	0.0000
DIST2(-1)	13614.22	2815.765	4.834997	0.0000
R-squared	0.600878	Mean dependent var		249446.6
Adjusted R-squared	0.574270	S.D. dependent var		53577.31
S.E. of regression	34958.12	Akaike info criterion		21.01032
Sum squared resid	3.67E+10	Schwarz criterion		21.14637
Log likelihood	-390.4953	F-statistic		22.58250
Durbin-Watson stat	1.964471	Prob(F-statistic)		0.000001

Figura 6 – Parâmetros do Modelo 5

Abrindo o escopo de estudo, como explicado precedentemente, permite-se, ao coeficiente R^2 , atingir o valor 60,1%. O erro padrão diminui para 14,1%.

Na representação gráfica do modelo 5 (gráfico 7), frente ao volume de vendas real, mostra-se que o modelo tem um poder explicativo das variações mensais um pouco melhor do que o modelo 4.

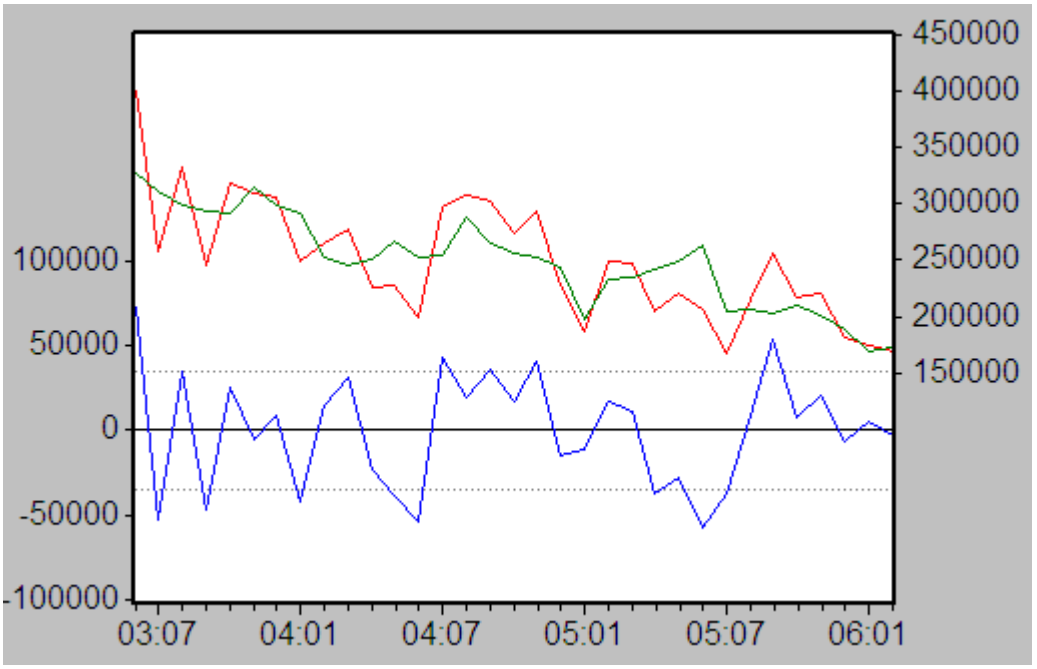


Gráfico 7 – Gráfico do Modelo 5

3.3.2.2.3 Modelo 6

Estando ainda insatisfeito com o poder explicativo do modelo 5, apresenta-se uma variante dele, chamada modelo 6, substituindo a variável DIST2(-1) pela variável DEP3(-1). Esta substituição obedece ao raciocínio seguinte: queria-se abrir o escopo do tipo de variável a serem levadas em conta no modelo 6. O modelo 5 utiliza dois tipos de variáveis: uma variável econômica IC1 e duas variáveis de logística DIST3 e DIST2(-1). No modelo 6 serão utilizadas três tipos de variáveis: uma variável econômica IC1, uma variável de logística DIST3 e uma variável de marketing DEP3(-1). Vale ressaltar que a variável econômica de índice de preço IC1 do produto P&G, contra o concorrente 1, traduz a comparação que o

consumidor faz no momento da compra do sabão em pó. O consumidor compara o preço do produto P&G com o do concorrente 1, escolhendo o produto que responde melhor as suas necessidades naquele momento. Assim, essa variável IC1 poderia, também, pertencer à categoria hábito do consumidor.

Os resultados do modelo 6 estão apresentados na figura 7 a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1515729.	315671.4	4.801603	0.0000
IC1	-10837.03	3168.477	-3.420266	0.0019
DIST3	-6542.707	977.6591	-6.692217	0.0000
DEP3(-1)	6318.741	3274.173	1.929874	0.0635

R-squared	0.633073	Mean dependent var	249446.6
Adjusted R-squared	0.595115	S.D. dependent var	53577.31
S.E. of regression	34091.54	Akaike info criterion	20.98682
Sum squared resid	3.37E+10	Schwarz criterion	21.16822
Log likelihood	-389.1075	F-statistic	16.67828
Durbin-Watson stat	2.336029	Prob(F-statistic)	0.000002

Figura 7 – Parâmetros do Modelo 6

O coeficiente R^2 do modelo 6 é de 63,3% e o erro padrão de 13,7%. Pode-se observar também que todos os valores de t dos coeficientes da regressão estão superiores ao valor critico $t_{crit} = 1{,}7$ com uma precisão de 95%.

A seguir, no gráfico 8, está apresentado o desempenho do modelo 6 de maneira gráfica:

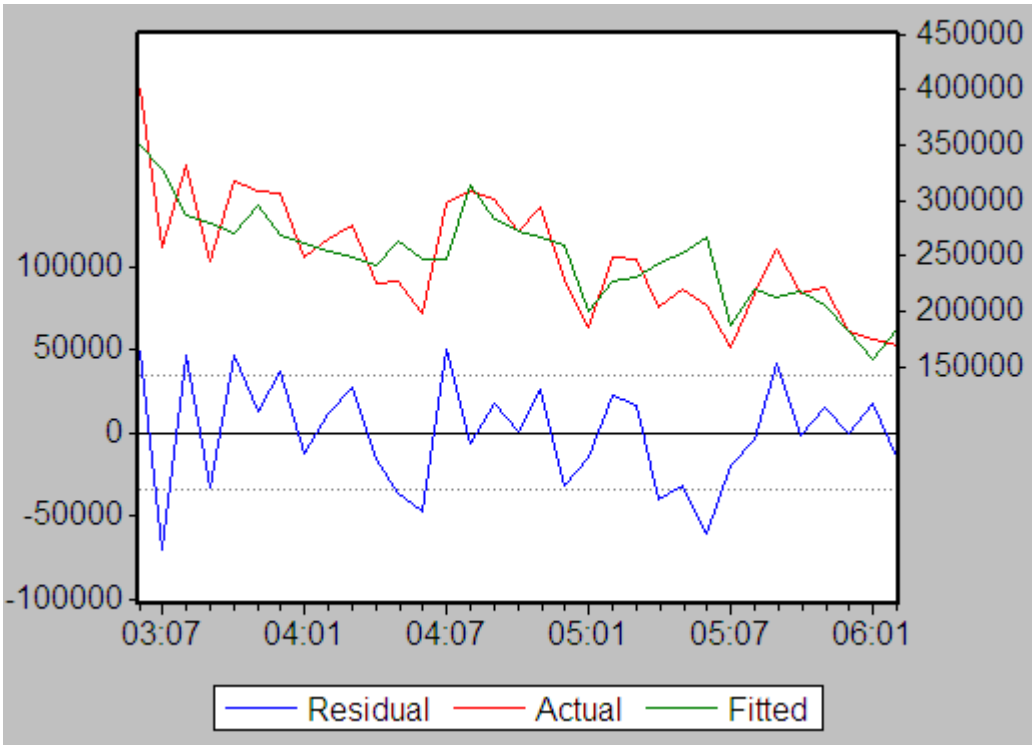


Gráfico 8 – Gráfico do Modelo 6

3.3.2.3 Modelos multiplicativos

3.3.2.3.1 Modelo 7

Neste parágrafo será apresentado um modelo um pouco diferente baseado sobre a teoria dos modelos lineares. Este modelo será chamado de modelo 7 e pertence a categoria dos modelos multiplicativos. Como explicado na revisão bibliográfica, podem-se usar modelos multiplicativos na teoria dos modelos multilineares através da linearização deles via o logaritmo neperiano. Essa técnica permite chegar-se a modelos com coeficiente R^2 mais altos neste trabalho.

O modelo 7 utiliza cinco variáveis, com três variáveis de preço dos concorrentes e duas variáveis de logística (LDIST3 e LPRE2(-1)). Os resultados desse modelo estão apresentados na figura 8 a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	73.45504	16.03653	4.580482	0.0002
LC1	4.840171	1.102128	4.391659	0.0003
LC2	-6.934619	1.388815	-4.993193	0.0001
LC3	-6.612196	1.318728	-5.014071	0.0001
LDIST3	-1.874449	0.239209	-7.836027	0.0000
LPRE2(-1)	-9.135677	3.352029	-2.725417	0.0130

R-squared	0.794947	Mean dependent var	12.34404
Adjusted R-squared	0.743684	S.D. dependent var	0.181823
S.E. of regression	0.092053	Akaike info criterion	-4.571617
Sum squared resid	0.169473	Schwarz criterion	-4.281287
Log likelihood	28.53862	F-statistic	15.50717
Durbin-Watson stat	2.021787	Prob(F-statistic)	0.000003

Figura 8 – Parâmetros do Modelo 7

Os resultados estão muito melhores do que nos modelos antigos. O R^2 é de 79,5% e todos os coeficientes t têm valores bem altos em valores absolutos.

Essa melhoria espetacular dos resultados tem várias fontes. A primeira fonte de melhoria é o uso de um modelo multiplicativo que ao longo dos testes em computador pareceu bem óbvia. Mas existe um outro fator que é o uso da variável de preço do concorrente 3. Essa variável tem uma leitura somente desde janeiro de 2004. Todas as outras variáveis têm uma leitura desde maio de 2003. Assim, ao usar a variável de preço do concorrente 3, o período de estudo se restringe ao período de janeiro de 2004 até fevereiro de 2006. Essa redução do período de análise do modelo pode ser uma fonte de explicação dos bons resultados do modelo 7, pois de maio de 2003 até agosto de 2003, o volume de vendas tem

variações grandes. Assim, essas oscilações grandes do volume de vendas não são levadas em conta e não influenciam os parâmetros do modelo.

Essa redução do período de estudo esta visível no gráfico 9 a seguir, assim como o bom desempenho do modelo 7:

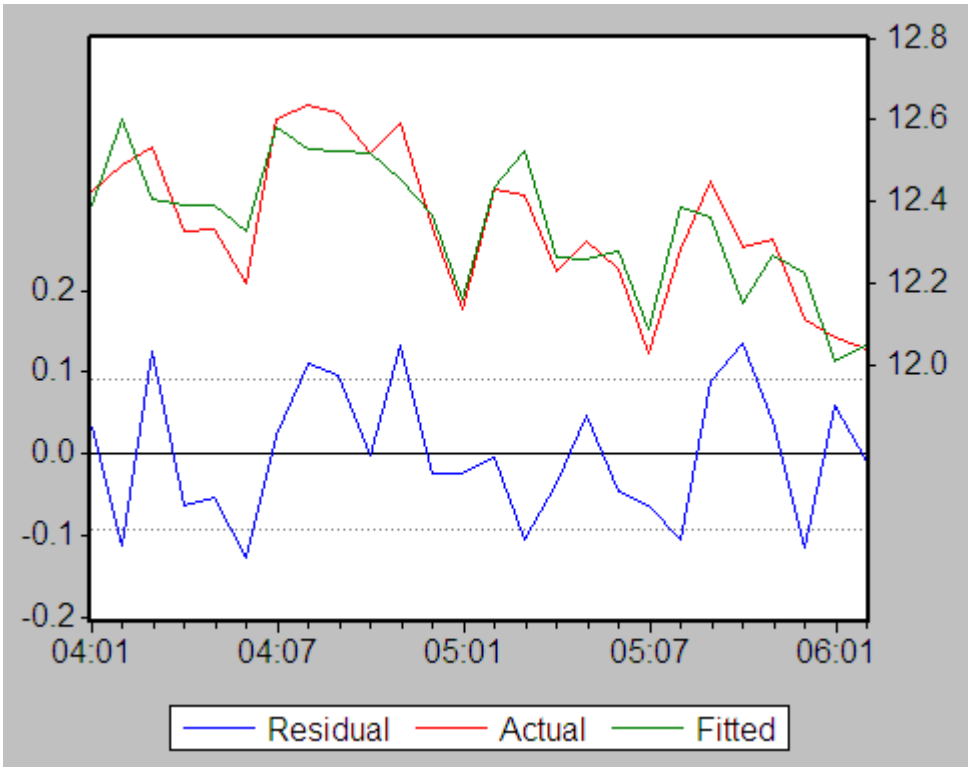


Gráfico 9 – Gráfico do Modelo 7

Apesar das excelentes qualidades deste modelo 7, deve-se reparar que nenhuma variável relativa ao produto P&G entra nele. Isso pode significar que o ambiente competitivo no qual o produto P&G atua tem mais influencia no seu volume de vendas do que nas variáveis do produto em si. O modelo 3 vem claramente se erguer contra essa hipótese, por estar constituído unicamente de variáveis referente ao produto P&G e chegar num valor de R^2 de 72,0%. Precisa-se parar neste modelo que apresenta resultados muito satisfatórios para tentar ver se faz sentido incluir uma ou algumas variáveis relativas ao produto P&G. Antes disso, deve-se assegurar que o modelo é satisfatório do ponto de vista matemático. Para tal

tarefa, faz-se uma análise de multicolinearidade do modelo 7, já que os testes t, F e de Durbin Watson estão todos verificados.

A seguir, no quadro 12, está apresentada a matriz de correlação do modelo 7.

	LC1	LC2	LC3	LDIST3	LPRE2(-1)	LVOL
LC1	1.000000	0.657917	-0.399072	0.381993	-0.143145	-0.082792
LC2	0.657917	1.000000	-0.414257	0.108456	-0.393269	-0.060322
LC3	-0.399072	-0.414257	1.000000	-0.819596	0.286290	0.325926
LDIST3	0.381993	0.108456	-0.819596	1.000000	-0.134421	-0.631804
LPRE2(-1)	-0.143145	-0.393269	0.286290	-0.134421	1.000000	-0.134772
LVOL	-0.082792	-0.060322	0.325926	-0.631804	-0.134772	1.000000

Quadro 12 – Matriz de correlação do Modelo 7

Pode-se observar que o maior coeficiente da matriz (em cinza) existe entre duas variáveis independentes do modelo, indicando a presença de multicolinearidade no modelo 7. Segundo Armstrong (1985), deve ser privilegiada a solução de combinar as duas variáveis incriminadas em vez de eliminá-las.

Após vários testes, cria-se uma nova variável a partir do logaritmo neperiano da distribuição do concorrente 3 LDIST3 e da variável de preço deste concorrente 3 (sem logaritmo neperiano) DC3, multiplicando as duas. A nova variável será chamada de MIX, com:

$$MIX = DC3 \times \ln(DIST3) = DC3 \times LDIST3 \qquad \textbf{(3.1)}$$

3.3.2.3.2 Modelo 8

O modelo 8, a seguir, é a evolução do modelo 7 sem as variáveis LC3 e LDIST3 e com a nova variável MIX, para tirar a multicolinearidade.

A figura 9, a seguir, oferece um resumo das principais características do modelo 8:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	61.27700	16.19507	3.783682	0.0011
LC1	4.962526	1.146175	4.329642	0.0003
LC2	-7.447238	1.379932	-5.396814	0.0000
LPRE2(-1)	-8.287763	3.402482	-2.435799	0.0239
MIX	-0.483654	0.058658	-8.245335	0.0000

R-squared	0.772075	Mean dependent var	12.34404
Adjusted R-squared	0.728661	S.D. dependent var	0.181823
S.E. of regression	0.094712	Akaike info criterion	-4.542790
Sum squared resid	0.188377	Schwarz criterion	-4.300849
Log likelihood	27.16387	F-statistic	17.78388
Durbin-Watson stat	2.115608	Prob(F-statistic)	0.000002

Figura 9 – Parâmetros do Modelo 8

O coeficiente R^2 diminui de 2,3% para chegar no valor de 77,2% no modelo 8, mas a maior parte da informação das variáveis LC3 e LDIST3 foi guardada. Tirar um ou outra variável do modelo significa perder mais de 10% no valor do coeficiente R^2. Assim verifica-se que o conselho de Armstrong (1985) é válido e bem útil.

O gráfico 10, a seguir, está apresentado o gráfico obtido com base no modelo 8, comparado com os valores reais do volume de vendas do produto P&G:

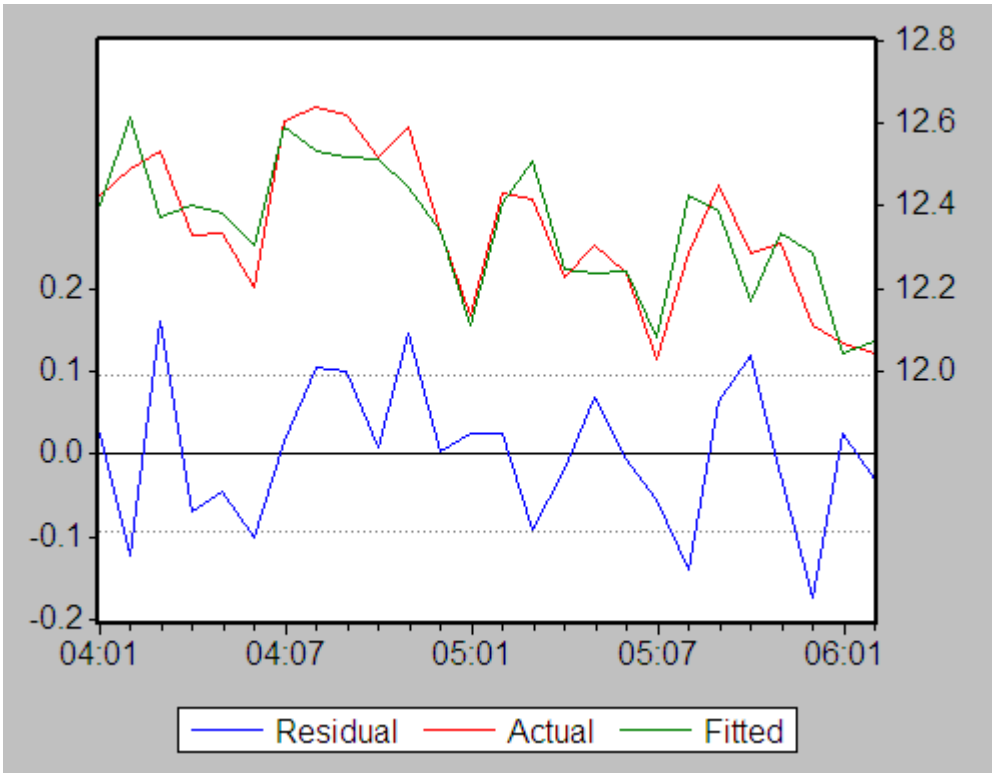


Gráfico 10 – Gráfico do Modelo 8

A seguir foi feita uma nova análise de multicolinearidade para verificar que depois da criação da variável MIX não existe mais prova de multicolinearidade.

	LC1	LC2	LPRE2(-1)	MIX	LVOL
LC1	1.000000	0.657917	-0.143145	0.175074	-0.082792
LC2	0.657917	1.000000	-0.393269	-0.325121	-0.060322
LPRE2(-1)	-0.143145	-0.393269	1.000000	0.132307	-0.134772
MIX	0.175074	-0.325121	0.132307	1.000000	-0.673081
LVOL	-0.082792	-0.060322	-0.134772	-0.673081	1.000000

Quadro 13 – Matriz de correlação do Modelo 8

Pode-se observar que o maior coeficiente da matriz de correlação (quadro 13) existe entre a variável independente MIX e a variável dependente LVOL. Isso significa que não existe evidência de multicolinearidade no modelo 8. O modelo 8 é, assim, validado matematicamente. **(3.2)**

A equação do modelo 8 se escreve da seguinte maneira:

$$DVOL = k \times DC1^{4,96} \times DC2^{-7,45} \times DPRE2(-1)^{-8,29} \times DIST3^{-0,48 \times DC3}$$

Com os coeficientes calculados pelo software E-Views reportados na tabela precedente. O coeficiente k é o exponential da constante C calculado pelo E-Views.

O próximo passo é a tentativa de incorporação de variáveis relativas ao produto P&G no modelo 8. Apesar do modelo 8 oferecer excelentes resultados, acredita-se que pelo menos uma variável relativa ao produto P&G deve entrar neste modelo.

3.3.2.3.3 Modelo 9

O modelo 9, a seguir, é o resultado de vários testes de incorporação das variáveis relativas ao produto P&G no modelo 8. A variável mais relevante, relativa ao produto P&G, a ser acrescentada no modelo 7, é a presença na loja, do produto P&G em logaritmo neperiano: LPRE. A figura 10, resumindo o desempenho do modelo 9, está apresentada a seguir:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	49.92040	17.83866	2.798440	0.0111
LC1	4.255353	1.232111	3.453710	0.0025
LC2	-6.112089	1.658302	-3.685752	0.0015
LPRE2(-1)	-7.833557	3.345995	-2.341174	0.0297
MIX	-0.495051	0.057992	-8.536531	0.0000
LPRE	1.887470	1.360426	1.387411	0.1806

R-squared	0.792086	Mean dependent var	12.34404
Adjusted R-squared	0.740107	S.D. dependent var	0.181823
S.E. of regression	0.092693	Akaike info criterion	-4.557758
Sum squared resid	0.171838	Schwarz criterion	-4.267428
Log likelihood	28.35846	F-statistic	15.23869
Durbin-Watson stat	2.050061	Prob(F-statistic)	0.000003

Figura 10 – Parâmetros do Modelo 9

O modelo 9 tem um desempenho similar ao modelo 8 com um coeficiente R^2 de 79,2%, uma estatística de Durbin Watson ligeiramente superior a 2 e uma probabilidade F muito pequena.

Apesar dessas características, a variável de preço do produto P&G tem um coeficiente t de 1,39 inferior ao limite de valor de t em valor absoluto para uma precisão de 95%. Isso significa que existe mais do que 5% de chance de que a variável LPRE seja irrelevante no modelo 9. Recalculando o valor limite de t, para uma precisão de 90%, verifica-se que o valor absoluto do coeficiente t da variável LPDV é superior a esse valor limite. Acredita-se, após os estudos feitos no começo desse capítulo sobre a relevância das variáveis do produto P&G nos modelos lineares, que a variável de presença na loja do produto P&G tem toda razão de figurar dentro das variáveis do modelo 9.

Assim, apesar de ter que abaixar a precisão do teste t para 90%, o que é um valor bastante aceitável, guarda-se a variável LPRE dentro do modelo 9.

Pode-se seguir o desempenho excepcional do modelo 9 através do gráfico 10 a seguir, comparando a curva real do logaritmo do volume de vendas com a curva fornecida pelo modelo.

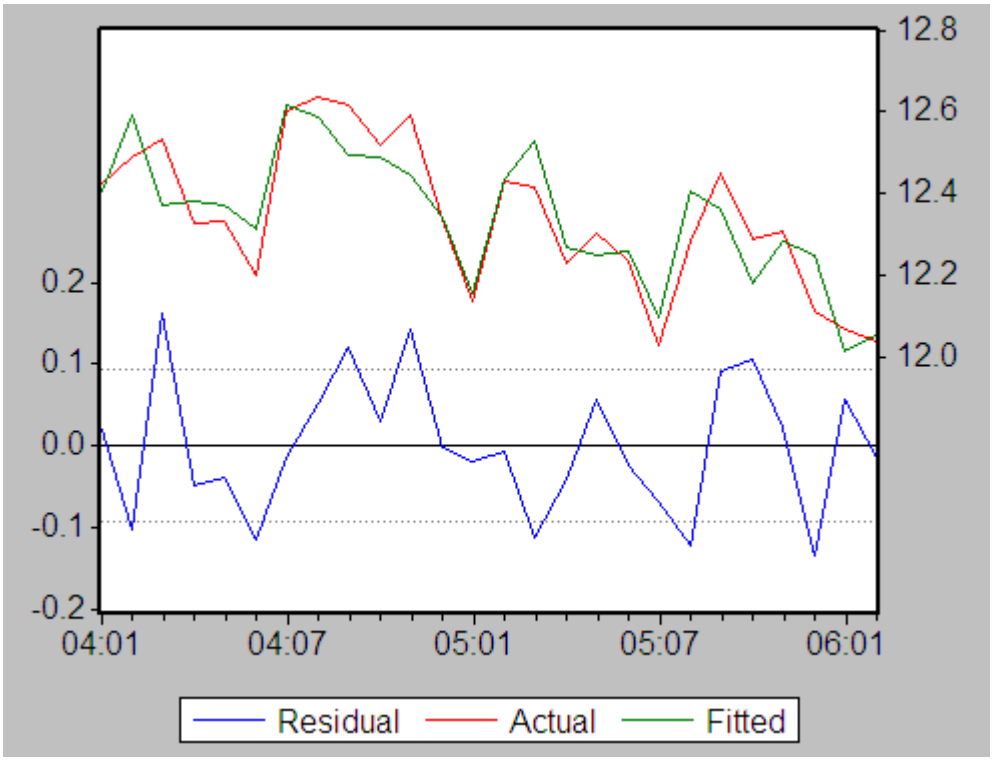


Gráfico 10 – Gráfico do Modelo 9

Acredita-se que este modelo constitui uma boa base para elaborar as previsões do volume de vendas e, assim, responder ao problema. Para validar o modelo 9, testa-se a sua multicolinearidade. A seguir, está apresentada a matriz de correlação do modelo 9 no quadro 14.

	LC1	LC2	LPRE2(-1)	MIX	LPRE	LVOL
LC1	1.000000	0.657917	-0.143145	0.175074	-0.038977	-0.082792
LC2	0.657917	1.000000	-0.393269	-0.325121	-0.559247	-0.060322
LPRE2(-1)	-0.143145	-0.393269	1.000000	0.132307	0.215455	-0.134772
MIX	0.175074	-0.325121	0.132307	1.000000	0.492090	-0.673081
LPRE	-0.038977	-0.559247	0.215455	0.492090	1.000000	0.034616
LVOL	-0.082792	-0.060322	-0.134772	-0.673081	0.034616	1.000000

Quadro 14 – Matriz de correlação do Modelo 9

Pode-se observar que o maior coeficiente da matriz está entre a variável independente MIX e a variável dependente LVOL, o que não evidencia nenhuma existência de multicolinearidade.

No capítulo a seguir, são comparados os melhores modelos multilineares com os métodos de série temporal utilizados atualmente pela empresa P&G. Os modelos multilineares julgados importante para este capitulo 4 são os modelos 3 e 9.

O modelo 3 foi retido porque ele leva em conta exclusivamente variáveis relativas ao produto P&G e oferece bons resultados. A equação 3.3 deste modelo é a seguinte:

$$DVOL = 80219{,}50 \times DPPG(-2) + 19173{,}86 \times DDIST(-2) + 11300{,}03 \times DPRE + 7042{,}269 \times DPDV - 11046{,}50 \times DEP(-2) - 5956{,}529 \times DEP - 2813084 \quad \textbf{(3.3)}$$

O modelo 9 é o de melhor desempenho e ele junta variáveis independentes de diferentes tipos como de preço, de logística, da concorrência etc. Por ser o modelo mais completo deste trabalho ele foi retido para ser comparado com os modelos de série temporal.

A equação do modelo 9 é apresentada a seguir:

$$DVOL = k \times DC1^{4,26} \times DC2^{-6,11} \times DPRE2(-1)^{-7,83} \times DIST3^{-0,49 \times DC3} \times DPRE^{1,89} \quad \textbf{(3.4)}$$

3.4 Resultados dos métodos de extrapolação

Nesta parte são implementados os principais métodos de série temporal, ou extrapolação, utilizados pela empresa P&G.

Os quatro principais modelos de extrapolação utilizados pela P&G são os seguintes:

- Método da média móvel
- Método da suavização exponencial
- Método de Holt
- Método de Winter

Como explicado na revisão bibliográfica, os dois primeiros métodos atendem à previsão de qualquer série de dados. Já o método de Holt é mais eficiente para prever serie de dados com tendência. O método de Winter apresenta melhores resultados para as series de dados com tendência e sazonalidade.

Assim, será estudada a série de dados do volume de vendas do produto P&G no mesmo período que os modelos de regressão multilineares: maio de 2003 até fevereiro de 2006.

3.4.1 Estudo da série temporal do volume de vendas

Para se fazer um estudo completo precisa-se estudar a presença ou não de tendência e de sazonalidade.

3.4.1.1 Tendência

A melhor maneira de evidenciar a tendência de um serie de dados é de traçar a curva assim como achar a tendência linear que melhor se aproxima dela.

O gráfico 11 a seguir mostra de maneira visual a presença de uma tendência na serie de dados de volume de vendas no período de maio 2003 até fevereiro 2006 estudado.

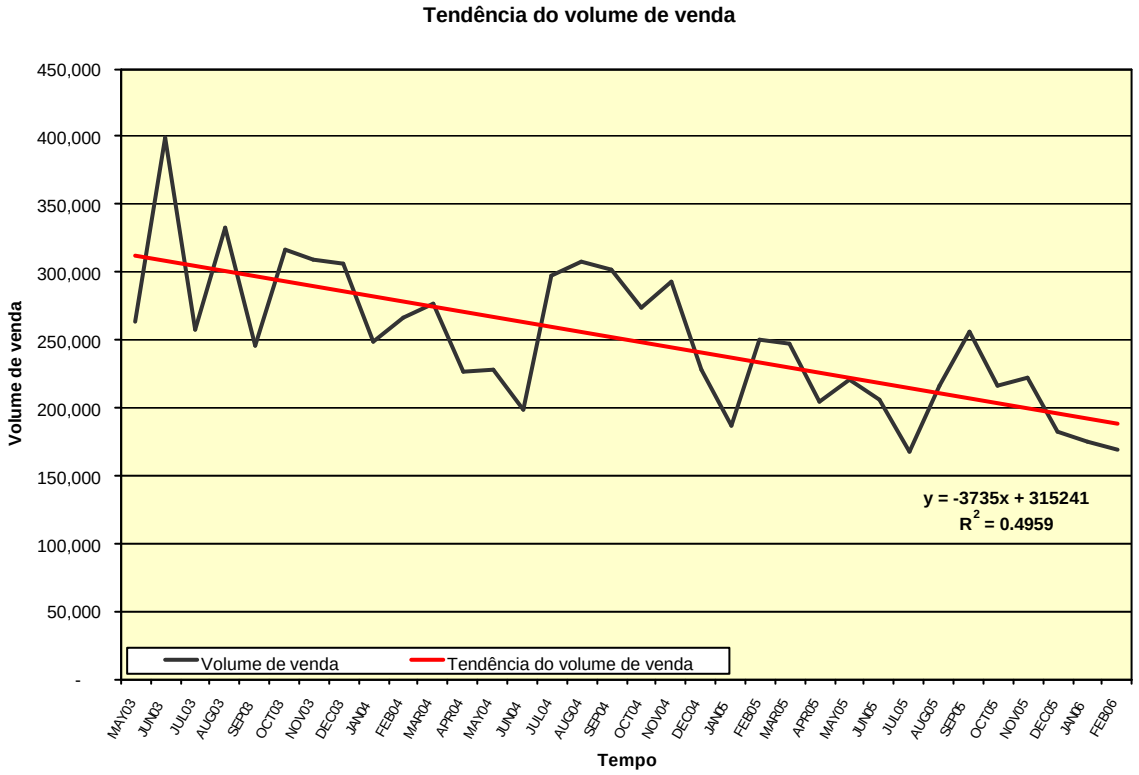


Gráfico 11 – Gráfico da tendência do volume de venda

A equação da reta de tendência é apresentada no gráfico e evidencia uma significativa tendência decrescente. O coeficiente R^2 é quase de 50%, o que mostra que esta tendência é pertinente para explicar a evolução do volume de vendas.

Uma vez a tendência evidenciada na série de dados do volume de vendas, fica necessário o desenvolvimento do método de Holt.

3.4.1.2 Sazonalidade

Para evidenciar a presença ou não de sazonalidade, precisa-se de mais do que um estudo gráfico. Necessita-se calcular os coeficiente de autocorrelação da serie de dados do volume de vendas para medir as relações entre a série de dados e ela mesma, defasada de alguns períodos.

De maneira intuitiva, poderia-se achar a presença de uma sazonalidade de período 12, já que os dados do volume de vendas são mensais. Assim para verificar esta hipótese precisa-se levar a análise de correlação com defasagens de mais de 12. O cálculo dos coeficientes de autocorrelação foi feito com a ajuda do software E-Views 2.0 e os resultados obtidos estão apresentados no correlograma a seguir (gráfico 12). O valor do coeficiente de correlação está apresentado em função do número de mês defasado. Pode-se observar uma clara tendência exponencial decrescente dos valores dos coeficientes de autocorrelação para se chegar num valor próximo de 0, depois de 10 períodos. Esta tendência apresentada em preto evidencia que nenhuma sazonalidade foi encontrada na série temporal do volume de vendas. Nenhum pico significativo está presente no valor 12 meses da defasagem: não existe sazonalidade anual.

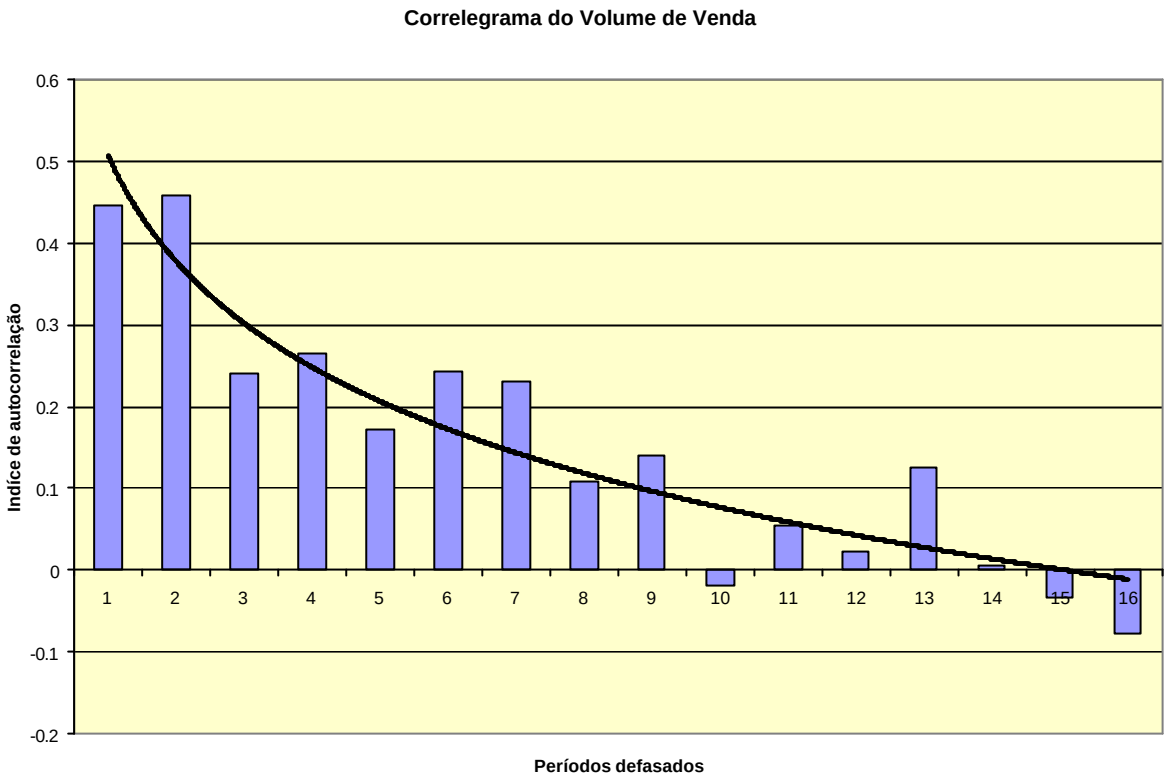


Gráfico 12 – Correlograma do volume de venda

O estudo foi feito até uma defasagem de 16 meses para poder verificar a presença ou não de sazonalidade anual. Não foram calculados coeficientes com defasagem maior do que 16 meses, pois a série temporal do volume de vendas apresenta 34 valores. Não seria relevante aumentar o número de períodos defasados, visto o resultado claro do estudo da autocorrelação.

A ausência de sazonalidade indica que a implementação do método de Winter é inútil, já que a única diferença com o de Holt é a aparição de um termo permitindo-se de levar em conta a sazonalidade dos dados na previsão.

3.4.2 Método da media móvel

O método mais simples a ser implementado é o da média móvel. O número de períodos a serem levados em conta na elaboração da média é escolhido de maneira a minimizar o MAE, como explicado na revisão bibliográfica. Este número, neste trabalho, é de 4 meses.

No gráfico 13, a seguir, está apresentado, de maneira gráfica, o desempenho do método de média móvel frente aos dados reais do volume de vendas.

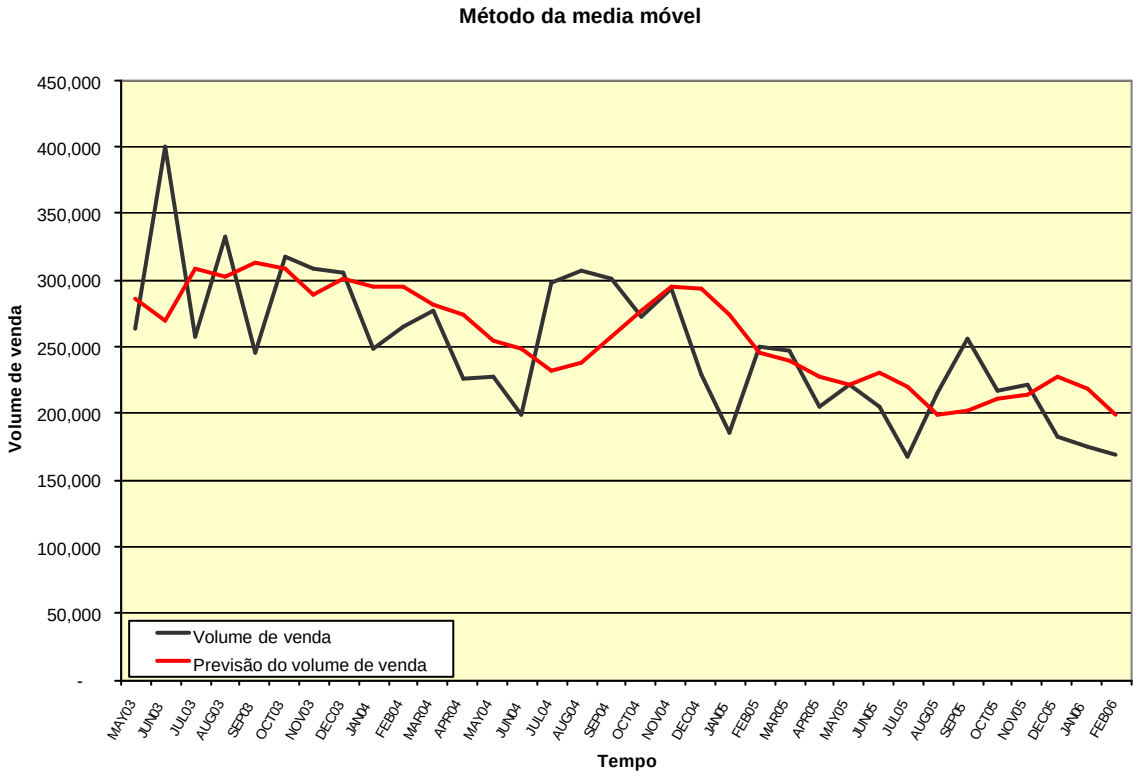


Gráfico 13 – Gráfico do método da media móvel

Pode-se observar que a resposta do modelo aos picos do volume de vendas é sempre atrasada, caracterizando bem os modelos de extrapolação que deduzem, do passado, o futuro.

O erro padrão deste modelo é 14,0%, o que é maior do que os modelos de regressão multilineares.

3.4.3 Método de suavização exponencial

Um método um pouco mais complexo é o de suavização exponencial. Como visto na revisão bibliográfica, precisa-se achar o fator de suavização a que minimize o MAE. O valor achado pelo solver do Excel é $a = 0{,}35$. Com este fator de suavização o erro padrão do modelo é 13,6%, superior aos erros padrões dos modelos multilineares apresentados no capítulo precedente. O gráfico 14, a seguir, compara os valores reais do volume de vendas com os achados pelo modelo de suavização exponential.

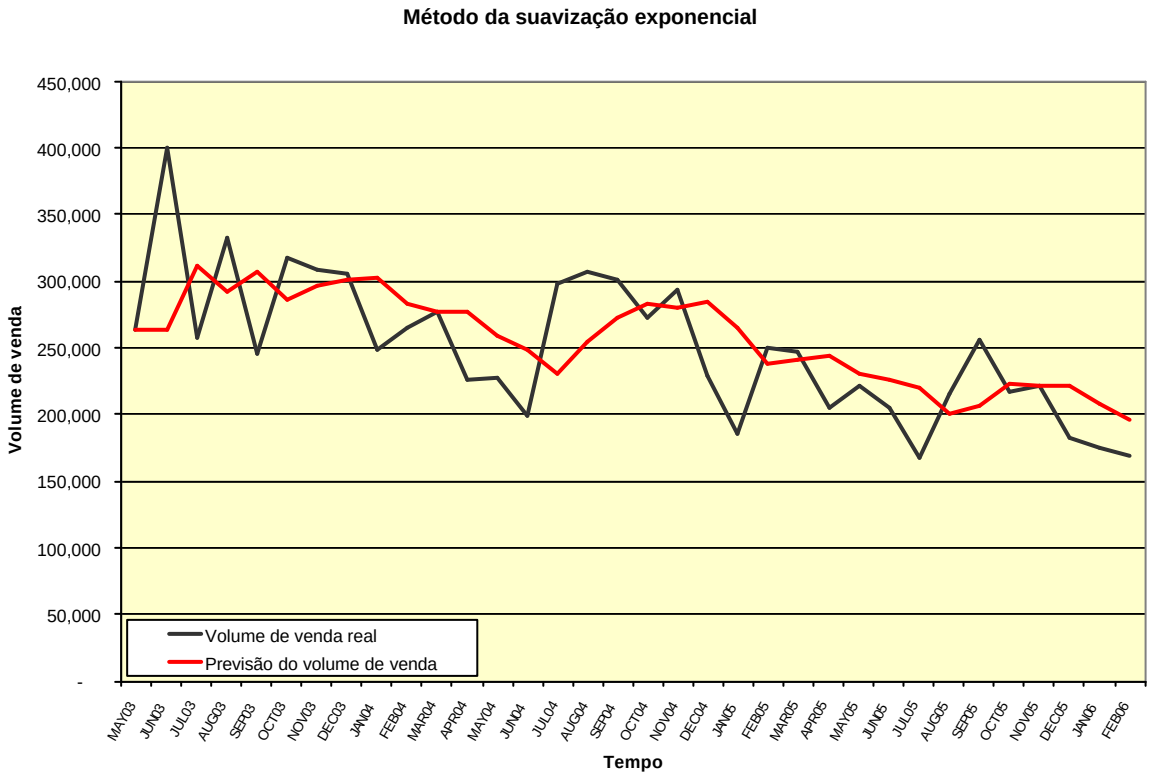


Gráfico 14 – Gráfico do método da suavização exponencial

3.4.4 Método de Holt

Como visto na revisão bibliográfica, o modelo de Holt se baseia no modelo de suavização exponencial, mas incorpora uma variável intermediaria de tendência para ter um poder explicativo maior. Precisam ser achados duas constantes a e b para o método de Holt. Os valores de a e b respondem ao mesmo critério de minimização do MAE e são respectivamente 0,42 e 0,11. O erro padrão deste modelo é 13,9%. Pode-se observar, no gráfico 15 a seguir, a resposta do modelo mais forte a cada pico do volume de vendas do que nos modelos de extrapolação precedentes.

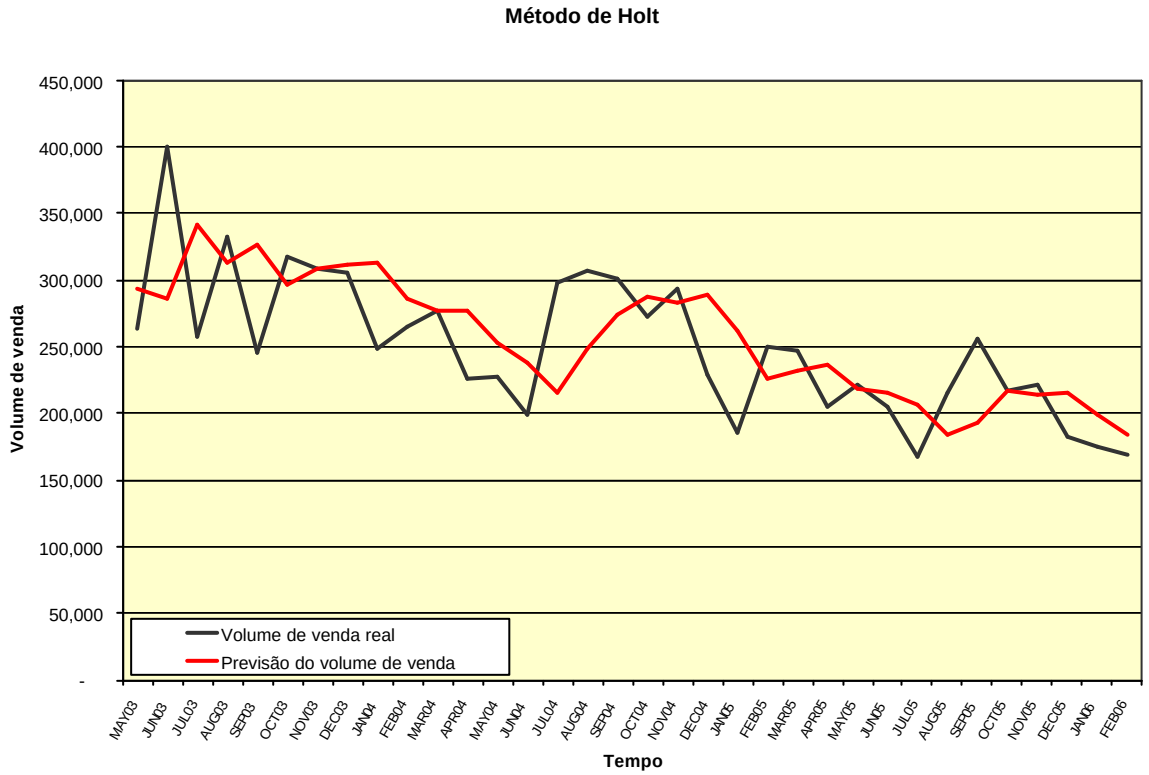


Gráfico 15 – Gráfico do método de Holt

4 Comparação dos métodos de previsão

Neste capítulo são confrontados os dois tipos de modelos expostos no trabalho: método de regressão multilinear e método de série temporal.

Esta comparação é composta de três partes. A primeira é uma comparação qualitativa entre os gráficos de volume de vendas obtidos através dos diferentes métodos e o gráfico de volume de vendas real. A segunda parte, consiste em uma comparação dos erros padrões calculados ao longo deste trabalho. Por fim, serão aplicados os modelos para prever os volumes de vendas dos meses de março e abril.

Os modelos escolhidos para realizar essas comparações de desempenho são:

- os modelos de extrapolação (média móvel, suavização exponencial e Holt) elaborados anteriormente no parágrafo 3.4.

- os modelos de regressão multilinear 3 e 9 como explicado no final do parágrafo 3.3.

4.1 Comparação qualitativa

Ao se comparar os gráficos 5 e 10 (modelos de regressão) com os gráficos 13, 14 e 15 (modelos de extrapolação) pode-se perceber, de maneira visual, que os modelos de regressão apresentam curvas que seguem muito melhor a curva de volume de vendas real do que os modelos de extrapolação.

Os modelos de extrapolação seguem bem a tendência do volume de vendas, mas reagem a uma mudança grande dele com um atraso. Isso vem da maneira como eles estão construídos: o valor futuro é construído baseado no valor passado e presente. Isto faz com que as curvas dos modelos sejam bem diferentes da curva de volume de vendas real.

Os modelos de extrapolação prevêem as mudanças grandes do volume de vendas no momento que elas acontecem. Estas mudanças grandes do volume de vendas têm uma ou várias causas que são capturadas pelas variáveis utilizadas nos modelos de regressão.

Assim, de uma maneira puramente qualitativa de observação dos gráficos dos modelos, pode-se ver o desempenho muito melhor dos modelos de regressão linear em relação aos modelos de extrapolação.

4.2 Comparação dos erros padrões

O critério de desempenho mais adequado para comparar os desempenho de dois modelos de regressão é o coeficiente R^2. Com base neste coeficiente, pode-se ver que o modelo 9 é mais preciso do que o 3. Para se comparar modelos de tipos diferentes, como neste caso, da comparação de modelos de regressão linear com modelos de extrapolação, precisa-se usar uma medida que permite esta comparação: o erro padrão.

O quadro 15, a seguir apresenta os valores dos erros padrões dos 5 modelos, assim como os coeficientes R^2 dos dois modelos de regressão linear:

	Método de extrapolação			Método de regressão linear	
	Média Móvel	Suavização exponencial	Método de Holt	Modelo 3 - variáveis relacionadas ao produto P&G	Modelo 9 - variáveis de todos os tipos
Erro Padrão	14.0%	13.6%	13.9%	11.3%	< 11.3%
Coeficiente R^2	-	-	-	71.9%	79.2%

Quadro 15 – Comparação do desempenho dos modelos

O erro padrão do modelo 9 não foi calculado porque ele se refere a um modelo em logaritmo, mas o valor seria menor do que o do modelo 3 porque o coeficiente R^2 é maior. O modelo 9 tem um poder explicativo maior do que o 3, então, o erro padrão é menor.

Observa-se que os métodos de extrapolação apresentam um erro padrão maior do que os modelos de regressão, o que é ligado às explicações do parágrafo precedente: os modelos

de regressão linear seguem melhor as grandes mudanças do volume de vendas, tendo, por conseqüência, um MAE menor, ou seja, um erro padrão menor (mesma média do volume de vendas para os dois tipos de modelos).

Assim, verifica-se, de maneira quantitativa, o melhor desempenho dos modelos de regressão linear frente aos modelos de extrapolação, no caso deste trabalho.

4.3 Aplicação dos modelos aos meses de março e abril

Os cincos modelos comparados neste quarto capítulo, são testados nos meses de março e abril de 2006 para se comparar os diferentes resultados e confrontá-los aos dados reais de volume de vendas levantados na empresa.

O gráfico a seguir (gráfico 16), mostra os volumes de vendas calculados por cada modelo, assim como os volumes de vendas que realmente ocorrem em março e abril:

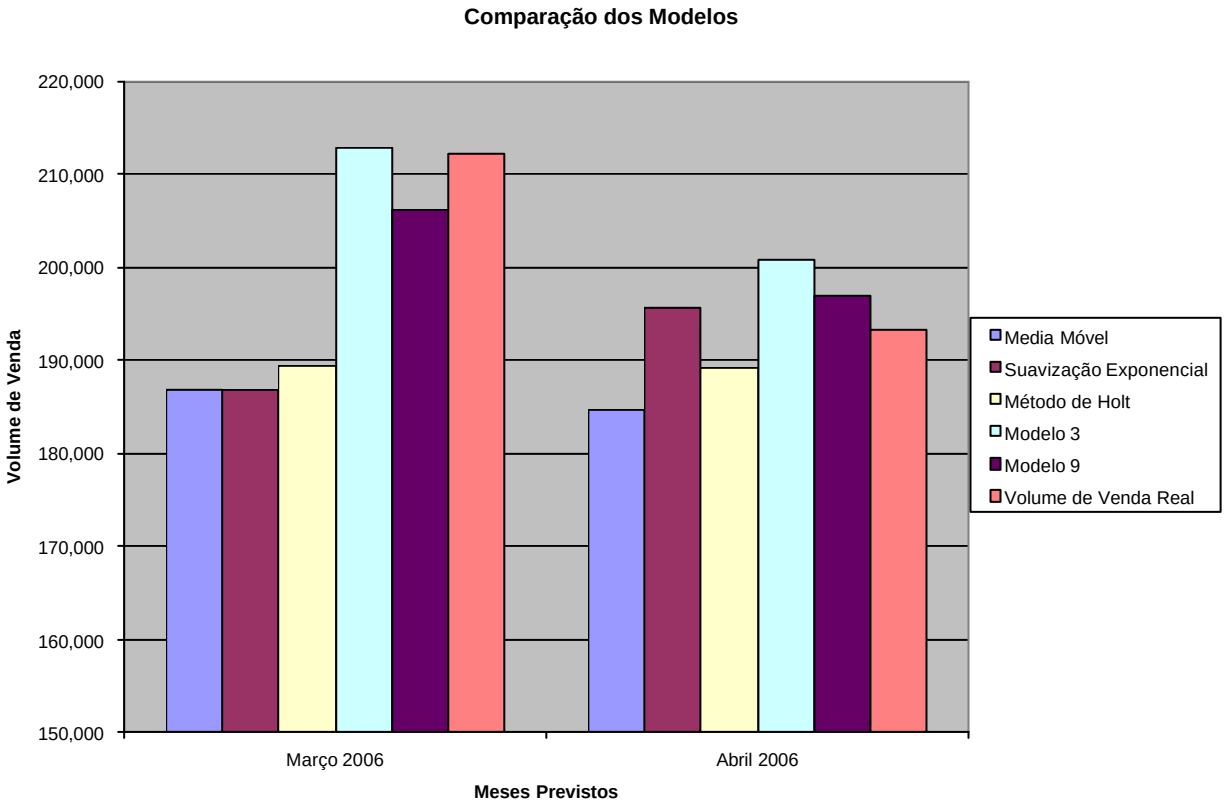


Gráfico 16 – Comparação dos modelos nos meses de março e abril de 2006

Observa-se que o modelo da média móvel tem o pior desempenho como constatado até agora. Pode-se ver que os métodos de extrapolação têm um desempenho bem inferior aos modelos de regressão linear para prever o mês de março. Isso vem da inversão de tendência do volume de vendas. O volume de vendas real volta a aumentar no mês de março e somente os modelos de regressão linear conseguem prever esta inversão de tendência. No mês de abril, eliminando o método da média móvel, é mais difícil determinar graficamente se os modelos de regressão têm um melhor desempenho.

Assim, são apresentados, no quadro 16 a seguir, os resultados de volume apresentados no gráfico 16, bem como o % de erro de cada modelo nos dois meses estudados.

	Metodo de extrapolação			Metodo de regressão linear		Realidade
	Media Móvel	Suavização exponencial	Metodo de Holt	Modelo 3 - variaveis relacionadas ao produto P&G	Modelo 9 - variaveis de todos os tipos	Volume de venda real
Março 2006	**186,925**	**186,784**	**189,428**	**212,782**	**206,173**	**212,332**
	12.0%	12.0%	10.8%	0.2%	2.9%	
Abril 2006	**184,624**	**195,712**	**189,168**	**200,926**	**197,040**	**193,324**
	4.5%	1.2%	2.1%	3.9%	1.9%	

Quadro 16 – Comparação dos modelos nos meses de março e abril de 2006

A luz das porcentagens de erro de cada modelo em abril, repara-se que os modelos de extrapolação têm um desempenho similar e até melhor do que os modelos de regressão linear.

Comparando os resultados dos modelos sobre dois meses não se permite tirar conclusões gerais indiscutíveis, mas confirma-se que os modelos de regressão linear reagem muito melhor a grandes mudanças ou inversão de tendência do volume de vendas do que os modelos de extrapolação. Analisando o quadro 16, confirma-se a melhor acurácia dos

modelos de regressão linear, validando a hipótese inicial de que o método de regressão linear permitiria atingir nível de precisão melhor do que os modelos atuais baseado no método de extrapolação.

5 Conclusões

A previsão do volume de vendas de um produto tem sido valorizada nas empresas por diversas razões. A primeira dela é a otimização da produção e do estoque, o que se traduze em uma redução do dinheiro parado e uma redução dos custos de armazenamento. Uma outra razão é a importância de antecipar os picos de volume, positivos para se ter um abastecimento correto das lojas e, assim, impedir uma ruptura na gôndola e negativos quando as suas vendas caiam. O caso da ruptura do produto na gôndola da loja é particularmente receado porque ele significa uma venda perdida, mas também uma venda sem esforços para a concorrência. No caso de uma previsão de queda do volume de vendas, podem ser estudados planos de ação para impedir que esta diminuição de volume de vendas aconteça.

A campo da previsão é vasto e podem ser achados métodos muito diferentes que correspondem a situações diferentes. O neófito tem um risco grande de não chegar a resultados satisfatórios se ele se concentra num método sem ter estudado um mínimo das características de cada método. O Esquema 2 classifica os métodos em quatro grandes grupos, que sempre se deve guardar em mente para saber quais tipos de resultados podem ser esperados. Uma vez escolhido o tipo de modelo mais adequado para o seu problema deve-se estudar em detalhe a teoria necessária a sua elaboração. Alguns métodos têm uma teoria matemática muito complexa que necessitam um investimento em tempo e energia consideráveis, como no caso dos métodos de regressão linear. Para as empresas, este tempo e esta complexidade se traduzem em investimento de dinheiro para alocar pessoas de competências suficientes para este tipo de trabalho. Assim, precisa-se avaliar o custo beneficio de tal método porque, após um investimento considerável, espera-se um resultado satisfatório, o que nem sempre acontece.

Em relação à resolução do problema deste trabalho, pode-se constatar, através das comparações efetuadas no capítulo quatro, que uma melhoria significativa da precisão foi alcançada. Esta melhoria da precisão se traduze por uma resposta instantânea dos modelos de regressão em relação às grandes mudanças do volume de vendas. Foi verificado que os modelos de extrapolação não têm esta capacidade de resposta, traduzindo-se por um erro padrão mais alto e uma precisão menor do que os modelos de regressão linear.

Assim, a proposta de melhorar o desempenho dos modelos de previsão da companhia, através do uso do método de regressão linear, foi realizada.

Em relação ao desenvolvimento do trabalho, foi verificado a grande diferença de complexidade entre os modelos de extrapolação e os modelos de regressão linear. Os métodos de extrapolação atingem resultados que satisfazem a maioria das exigências com um investimento razoável de tempo e energia. A relação custo beneficio de tais modelos é muito atraente e explica por que muitas empresas os usam. Os modelos de regressão linear devem ser usados somente em caso de necessidade de uma precisão maior, muitas vezes exigida pelo ambiente competitivo do mercado. Eles implicam um investimento em tempo e um conhecimento muito além dos modelos de extrapolação. A grande complexidade destes modelos tem como conseqüência a incerteza do resultado. Não se sabe, até último momento, se o resultado realmente será melhor do que um método mais simples, como o de extrapolação. Somente uma boa estruturação do problema, das variáveis e do raciocínio permitem chegar a melhorias significativas. Ressalta-se que, uma vez o modelo de regressão linear implementado, são necessárias as previsões das variáveis independentes antes de aplicar o modelo para se chegar na previsão da variável dependente estudada.

Assim, este grande trabalho na coleta das variáveis, na sua escolha para se construir o modelo, na validação matemática do modelo, agregado ao trabalho necessário para se realizar

a previsão propriamente dita, pode desmotivar muitas pessoas em busca de um modelo de previsão.

Deixando de lado o aspecto de desempenho dos modelos, o método de regressão linear proporciona, através da sua realização, um grande conhecimento do ramo onde se atua. No caso da busca de solução ao problema deste trabalho, foi possível identificar quais são os reais concorrentes do produto P&G, quais características dos produtos concorrentes têm mais impacto nas vendas do produto P&G e até quais características do produto P&G têm mais impacto nas suas vendas. Por exemplo, foi observado que a distribuição é uma variável fundamental para explicar o volume de vendas. Estas informações têm um valor muito grande para a empresa, que não pode ser medido. Assim, apesar da incerteza de resultado, a pessoa que realizará o trabalho de regressão linear terá um conhecimento do ambiente competitivo e dos produtos muito grande, o que não seria possível atingir somente aplicando o método de extrapolação.

Referências*

* De acordo com:
ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 6023:** informação e documentação: referências: elaboração. Rio de Janeiro, 2002.

- ALVIM, André Chang. **Previsão de demanda no varejo.** 2003. 93f. Trabalho de conclusão de curso (Trabalho de formatura) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2003.

- ARMSTRONG J. Scott. **Long-range forecasting from crystal ball to computer.** 2. ed. New York: Wiley, 1985.

- ARMSTRONG J. Scott. **Principles of forecasting, a handbook for researchers and practitioners.** 2. ed. Boston : Kluwer Academic Publishers, 2002.

- GUJARATI, Damodar N. **Econometria básica.** 3. ed. São Paulo: Makron Books, 2005.

- JARRET, Jeffrey. **Business forecasting methods.** Oxford: Basil Blackwell Ltd, 1987.

- MAKRIDAKIS, Spyros; WHEELWRIGHT, Steven; HYNDMAN, Rob. **Forecasting: method and applications.** 3. ed. New York: John Wiley & Sons, 1998.

- MOHALLEM, Thiago Pereira. **Previsão de vendas de um produto através de modelagem econométrica.** 2003. Trabalho de graduação – Instituto Tecnológico da Aeronáutica. São José dos Campos, 2003.

- NASCIMENTO, Gabriel Rotolo. **Previsão de preços do Mercado sucro – alcooleiro utilizando redes neurais.** 2004. 100f. Trabalho de conclusão de curso (Trabalho de formatura) – Escola Politécnica, Universidade de São Paulo. São Paulo, 2004.

- PINDYCK, Robert S.; RUBINFELD, Daniel L. **Econometric models and economic forecasts.** 3. ed. New York; McGraw-Hill, 1991.

Apêndices

Apêndice A – Variáveis referentes aos produtos estudados

	Produto P&G					
	Volume de Venda	**Preço**	**Distribuição**	**Presença**	**PDV**	**PEA**
Maio´03	264,096	5.41	89	94	4	9
Jun´03	399,711	5.41	89	94	4	9
Jul´03	257,887	5.33	90	95	7	14
Ago´03	332,510	5.35	90	95	7	14
Set´03	245,885	5.33	90	95	6	11
Out´03	317,222	5.29	90	95	6	11
Nov´03	308,541	5.00	90	95	3	6
Dez´03	305,782	5.07	90	95	3	6
Jan´04	248,835	5.15	89	92	5	13
Fev´04	265,683	5.26	89	92	5	13
Mar´04	277,212	5.06	88	95	4	12
Abr´04	226,395	5.00	88	95	4	12
Maio´04	227,430	4.97	89	96	3	9
Jun´04	198,995	4.99	89	96	3	9
Jul´04	297,617	5.01	87	98	4	10
Ago´04	307,790	4.74	87	98	4	10
Set´04	301,659	4.99	88	95	11	13
Out´04	273,323	5.09	88	95	11	13
Nov´04	293,489	5.06	86	96	7	14
Dez´04	228,539	5.04	86	96	7	14
Jan´05	186,281	5.17	84	98	8	11
Fev´05	249,697	5.08	84	98	8	11
Mar´05	246,873	5.03	85	98	6	12
Abr´05	204,339	5.01	85	98	6	12
Maio´05	220,862	4.96	86	98	9	15
Jun´05	205,743	4.95	86	98	9	15
Jul´05	167,328	4.97	86	97	9	14
Ago´05	216,034	4.85	86	97	9	14
Set´05	255,651	4.87	85	97	10	12
Out´05	216,725	4.86	85	97	10	12
Nov´05	221,534	4.68	85	94	9	12
Dez´05	182,179	4.73	85	94	9	12
Jan´06	174,395	4.74	81	94	7	9
Fev´06	169,592	4.68	81	94	7	9

	Concorrente 1				
	Preço	Distribuição	Presença	PDV	PEA
Maio´03	6.38	99	98	7	25
Jun´03	6.18	99	98	7	25
Jul´03	6.04	100	99	11	31
Ago´03	5.95	100	99	11	31
Set´03	6.01	99	98	13	34
Out´03	5.95	99	98	13	34
Nov´03	6.03	99	98	10	22
Dez´03	5.98	99	98	10	22
Jan´04	5.99	99	98	7	29
Fev´04	5.90	99	98	7	29
Mar´04	5.51	99	98	7	30
Abr´04	5.47	99	98	7	30
Maio´04	5.48	99	98	5	24
Jun´04	5.40	99	98	5	24
Jul´04	5.51	98	98	5	23
Ago´04	5.54	98	98	5	23
Set´04	5.70	98	97	9	21
Out´04	5.74	98	97	9	21
Nov´04	5.76	98	97	8	22
Dez´04	5.77	98	97	8	22
Jan´05	5.72	98	97	7	19
Fev´05	5.77	98	97	7	19
Mar´05	5.77	98	98	11	26
Abr´05	5.79	98	98	11	26
Maio´05	5.87	99	98	8	28
Jun´05	5.87	99	98	8	28
Jul´05	5.84	99	98	8	34
Ago´05	5.82	99	98	8	34
Set´05	5.86	99	97	8	29
Out´05	5.79	99	97	8	29
Nov´05	5.63	99	96	7	28
Dez´05	5.69	99	96	7	28
Jan´06	5.63	99	97	7	23
Fev´06	5.70	99	97	7	23

Quadro 18 – Variáveis referentes ao produto do concorrente 1

	Concorrente 2				
	Preço	Distribuição	Presença	PDV	PEA
Maio´03	5.39	94	98	5	9
Jun´03	5.31	94	98	5	9
Jul´03	5.15	94	98	6	12
Ago´03	5.13	94	98	6	12
Set´03	5.08	94	97	10	13
Out´03	5.03	94	97	10	13
Nov´03	5.06	94	97	6	10
Dez´03	5.15	94	97	6	10
Jan´04	5.17	93	96	2	11
Fev´04	5.04	93	96	2	11
Mar´04	4.74	93	98	4	11
Abr´04	4.63	93	98	4	11
Maio´04	4.61	93	97	7	9
Jun´04	4.67	93	97	7	9
Jul´04	4.62	92	97	4	8
Ago´04	4.72	92	97	4	8
Set´04	4.75	92	97	6	16
Out´04	4.76	92	97	6	16
Nov´04	4.79	91	97	7	15
Dez´04	4.81	91	97	7	15
Jan´05	4.80	92	96	9	10
Fev´05	4.79	92	96	9	10
Mar´05	4.74	92	98	10	16
Abr´05	4.74	92	98	10	16
Maio´05	4.74	93	97	10	12
Jun´05	4.77	93	97	10	12
Jul´05	4.83	92	97	9	11
Ago´05	4.69	92	97	9	11
Set´05	4.68	93	97	8	12
Out´05	4.80	93	97	8	12
Nov´05	4.70	92	97	5	7
Dez´05	4.77	92	97	5	7
Jan´06	4.78	91	96	4	8
Fev´06	4.87	91	96	4	8

Quadro 19 – Variáveis referentes ao produto do concorrente 2

	Concorrente 3			
	Preço	**Distribuição**	**Presença**	**PEA**
Maio´03	-	41	92	8
Jun´03	-	41	92	8
Jul´03	-	43	94	4
Ago´03	-	43	94	4
Set´03	-	46	95	3
Out´03	-	46	95	3
Nov´03	-	52	94	2
Dez´03	-	52	94	2
Jan´04	4.00	51	93	6
Fev´04	3.99	51	93	6
Mar´04	4.27	47	93	4
Abr´04	4.24	47	93	4
Maio´04	4.32	45	95	4
Jun´04	4.32	45	95	4
Jul´04	4.21	47	96	5
Ago´04	4.18	47	96	5
Set´04	4.21	48	97	5
Out´04	4.22	48	97	5
Nov´04	4.20	50	97	3
Dez´04	4.24	50	97	3
Jan´05	4.26	54	94	3
Fev´05	4.19	54	94	3
Mar´05	4.15	55	95	4
Abr´05	4.21	55	95	4
Maio´05	4.21	57	92	6
Jun´05	4.22	57	92	6
Jul´05	4.07	68	93	8
Ago´05	4.00	68	93	8
Set´05	4.03	69	94	10
Out´05	4.02	69	94	10
Nov´05	3.91	72	91	6
Dez´05	3.91	72	91	6
Jan´06	3.96	74	93	7
Fev´06	3.95	74	93	7

Quadro 20 – Variáveis referentes ao produto do concorrente 3

Apêndice B – logaritmo neperiano das variáveis

Variável (LN)	Símbolo
Volume de Venda produto P&G	LVOL
Preço produto P&G	LPPG
Distribuição produto P&G	LDIST
Presença na Loja produto P&G	LPRE
PDV produto P&G	LPDV
Pontos extras produto P&G	LEP
Preço Concorrente 1	LC1
Preço Concorrente 2	LC2
Preço Concorrente 3	LC3
Preço ponderado 3 concorrentes	LP3
Preço ponderado 2 concorrentes	LP2
Índice preço concorrente 1	LIC1
Índice preço concorrente 2	LIC2
Índice preço concorrente 3	LIC3
Índice preço ponderado 3 concorrentes	LIP3
Índice preço ponderado 2 concorrentes	LIP2
Distribuição concorrente 1	LDIST1
Presença na loja concorrente 1	LPRE1
PDV concorrente 1	LPDV1
Pontos Extras concorrente 1	LEP1
Distribuição concorrente 2	LDIST2
Presença na loja concorrente 2	LPRE2
PDV concorrente 2	LPDV2
Pontos Extras concorrente 2	LEP2
Distribuição concorrente 3	LDIST3
Presença na loja concorrente 3	LPRE3
Pontos Extras concorrente 3	LEP3

Quadro 21 – Símbolos das variáveis em logaritmo