

ÉRIKA MONIQUE FERREIRA JANNUZZELLI

Aplicação do modelo APT para definição do retorno de mercado: uma aplicação no índice IBOVESPA

ÉRIKA MONIQUE FERREIRA JANNUZZELLI

Aplicação do modelo APT para definição do retorno de mercado: uma aplicação no índice IBOVESPA

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia Mecânica da Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Mecânica.

Orientador (a): Prof. Dr. Francisco Alexandre de Oliveira

Guaratinguetá - SP
2018

J34a	Jannuzzelli, Érika Monique Ferreira Aplicação do modelo APT para definição do retorno de mercado: uma aplicação no índice IBOVESPA. / Érika Monique Ferreira Jannuzzelli– Guaratinguetá, 2018. 52 f : il. Bibliografia: f. 50-52 Trabalho de Graduação em Engenharia Mecânica – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2018. Orientador: Prof. Dr. Francisco Alexandre de Oliveira 1. Mercado de capitais 2. Análise de regressão 3. Desenvolvimento econômico I. Título CDU 339.727
------	---

Ana Cristina Figueiredo Loureiro
Bibliotecária CRB-8/7094

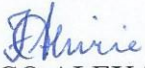
ÉRIKA MONIQUE FERREIRA JANNUZZELLI

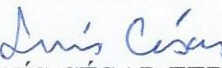
ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO
PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE
"GRADUADO EM ENGENHARIA MECÂNICA"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE
GRADUAÇÃO EM ENGENHARIA MECÂNICA

Prof. Dr. MARCELO SAMPAIO MARTINS
Coordenador

BANCA EXAMINADORA:


Prof. Dr. FRANCISCO ALEXANDRE DE OLIVEIRA
Orientador/UNESP-FEG


Prof. Ms. LUÍS CÉSAR FERREIRA MOTA BARBOSA
UNESP-FEG


Prof. Ms. SÉRGIO TENÓRIO DOS SANTOS NETO
FATEC-Guaratinguetá

Novembro de 2018

AGRADECIMENTOS

Em primeiro lugar agradeço a Deus, fonte da vida e da graça. Agradeço pela minha vida, minha inteligência, minha família e meus amigos,

ao meu orientador, *Prof. Dr. Francisco Alexandre de Oliveira* que jamais deixou de me incentivar. Sem a sua orientação, dedicação e auxílio, o estudo aqui apresentado seria praticamente impossível.

À minha mãe *Carmen Silvia Ferreira*, que apesar das dificuldades enfrentadas, sempre incentivou meus estudos.

às funcionárias da Biblioteca do Campus de Guaratinguetá pela dedicação, presteza e principalmente pela vontade de ajudar,

aos funcionários da Faculdade de Engenharia do Campos de Guaratinguetá pela dedicação e alegria no atendimento.

“Só sabemos com exatidão quando sabemos pouco;
à medida que vamos adquirindo conhecimento,
instala-se a dúvida.”

Goethe

RESUMO

O presente trabalho teve como objetivo aplicar a Teoria de Precificação por Arbitragem – *Arbitrage Pricing Theory* (APT) para determinar o índice IBOVESPA, após a crise financeira, por meio do uso da regressão multivariada. O método de pesquisa utilizado foi modelagem e simulação, analisando se o modelo encontrado se adequa aos índices de desempenho de um modelo de regressão, nesse caso os índices R^2 , R^2 ajustado e R^2 previsão. A coleta de dados para modelagem foi feita considerando a necessidade de se incluir mais variáveis para adequação do modelo, não sendo feita de uma única vez. Deste modo, na última análise a variável *Volume médio diário negociado na Bovespa* melhorou os índices do modelo, sendo incluída. Além disso, o modelo trabalhado é obtido com a transformação linear por logaritmo natural, o que indica que a estimativa do IBOVESPA, considerando a análise de regressão, não é totalmente linear, devendo ser feita uma linearização para depois ser aplicada a regressão linear. O modelo que melhor se ajustou foi o modelo logarítmico natural, considerando as três variáveis estudadas, que foram o IPCA, o Câmbio e a variável *Volume médio diário negociado na Bovespa*, a qual define toda a dinâmica de mercado.

PALAVRAS-CHAVE: Teoria de Precificação por Arbitragem. IBOVESPA. Regressão Multivariada. Variáveis de Mercado. Índices de desempenho.

ABSTRACT

The objective of this study was to apply the Arbitrage Pricing Theory (APT) to determine the IBOVESPA index, after the financial crisis, through the use of multivariate regression. The research method used was modeling and simulation, analyzing if the model found is adequate to the performance indexes of a regression model, in this case the indexes R^2 , R^2 adjusted and R^2 forecast. Data collection for modeling was done considering the need to include more variables to fit the model, not being done in a single time. Thus, in the last analysis the variable Average daily trading volume on the Bovespa improved the indexes of the model, being included. In addition, the worked model is obtained with the linear transformation by natural logarithm, which indicates that the IBOVESPA estimate, considering the regression analysis, is not totally linear, and a linearization must be done to be applied to the linear regression. The model that best fit was the natural logarithmic model, considering the three variables studied, which were the IPCA, the Exchange rate and the average daily volume traded on the Bovespa, which defines all the market dynamics.

KEYWORDS: Arbitrage Pricing Theory. IBOVESPA. Multivariate Regression. Marketplace variables. *Volume médio diário negociado na Bovespa* variable. Performance indexes.

LISTA DE ILUSTRAÇÕES

Figura 1 – Classificação da Pesquisa.....	13
Figura 2 – Abordagem metodológica desta Pesquisa.....	14
Figura 3 – Gráfico Risco total <i>versus</i> Número de títulos na carteira	17
Figura 4 – Distribuições de probabilidades conjuntas e o sinal da covariância entre X e Y	22
Figura 5 – Um exemplo de extrapolação em regressão múltipla	39
Figura 6 – Boxplot do Retorno IBOVESPA	43
Figura 7 – Análise da variável <i>Volume médio diário negociado no Bovespa</i>	45
Gráfico 1 – Probabilidade Normal.....	47
Gráfico 2 – Resíduos versus Ordem.....	47
Gráfico 3 – Resíduos versus Valores ajustados.....	48

LISTA DE TABELAS

Tabela 1 – Dados para Regressão Linear Múltipla.....	26
Tabela 2 – Análise de Variância para Testar a Significância na Regressão Múltipla	32
Tabela 3 – Relação entre os Artigos selecionados e os fatores macroeconômicos abordados em cada um.....	40
Tabela 4 – Resultados colhidos em cada etapa de análise Minitab	44
Tabela 5 – Resultados de R^2 colhidos no modelo final de análise Minitab.....	45
Tabela 6 – Resultados colhidos no modelo final de análise Minitab	46

LISTA DE ABREVIATURAS E SIGLAS

APT	Arbitrage Pricing Theory
IBOVESPA	Índice da Bolsa de Valores de São Paulo
CAPM	Capital Asset Pricing Theory
PNB	Produto Interno Bruto
IPCA	Índice de Preços ao Consumidor
IBGE	Instituto Brasileiro de Geografia e Estatística
PIB	Produto Interno Bruto
SELIC	Sistema Especial de Liquidação e Custódia
NYSE	New York Stock Exchange
VIF	Variance inflation factors
VMDNB	Volume médio diário negociado na Bovespa

LISTA DE SÍMBOLOS

β	regressor
R	retorno total efetivo do mês
$\bar{R}$	parte esperada do retorno
U	parcela inesperada do retorno
ε	risco não sistemático
m	risco sistemático
σ_{XY}	covariância
ρ_{XY}	correlação
R^2	coeficiente de determinação múltipla

SUMÁRIO

1	INTRODUÇÃO	12
1.1	OBJETIVOS	12
1.1.1	Geral	12
1.1.2	Específicos	12
1.2	JUSTIFICATIVA	12
1.3	MÉTODO DE PESQUISA	13
1.4	ESTRUTURA DO TRABALHO	14
2	A TEORIA DE PRECIFICAÇÃO POR ARBITRAGEM – APT	15
2.1	CONSIDERAÇÕES INICIAIS	15
2.2	APT X CAPM	18
2.3	DEFINIÇÃO DAS VARIÁVEIS OU FATORES EM UM MODELO APT	20
2.4	MEDIDAS DE DEPENDÊNCIA ENTRE VARIÁVEIS	20
3	MODELO DE REGRESSÃO MÚLTIPLA	25
3.1	CONSIDERAÇÕES INICIAIS	25
3.2	ANÁLISE DE REGRESSÃO	25
3.3	MEDIDAS DE DESEMPENHO DE UM MODELO DE REGRESSÃO	29
3.4	ETAPAS NO DESENVOLVIMENTO DE UM MODELO DE PREVISÃO APLICANDO REGRESSÃO MÚLTIPLA	37
4	APLICAÇÃO DA REGRESSÃO MÚLTIPLA AO APT	40
4.1	CONSIDERAÇÕES INICIAIS	40
4.2	ESPECIFICAÇÃO DOS FATORES	40
4.3	COLETA DE DADOS	41
4.4	APLICAÇÃO DA REGRESSÃO MÚLTIPLA NA ESTIMAÇÃO DE UM MODELO APT	42
4.5	ANÁLISE DOS RESULTADOS	44
5	CONCLUSÃO	49
	REFERÊNCIAS BIBLIOGRÁFICAS	50

1 INTRODUÇÃO

Existem diversos métodos para se estimar o retorno dos ativos de capitais e esses métodos podem ser aplicados às ações particulares ou para definir retornos de mercado. Nesse caso, o estudo será aplicado para o retorno de mercado no setor brasileiro especificamente no índice IBOVESPA.

Um dos métodos utilizados é o APT, *Arbitrage Pricing Theory*, Teoria de Precificação por Arbitragem, que tenta prever o retorno de um ativo, baseado em fatores macroeconômicos, tais como o PIB, taxa de juros, taxa de inflação, entre outros. Ele veio para confrontar a teoria mais utilizada até hoje, o Modelo de Precificação de ativos, *Capital Asset Pricing Model*, CAPM. O CAPM tem algumas deficiências, não especifica os fatores, se utiliza apenas um regressor e esse beta é calculado apenas pela covariância entre o ativo e o mercado sobre a variância de mercado. Já o APT diz que os retornos são gerados por uma série de fatores macroeconômicos, havendo assim mais de um beta no modelo, então ele pode se aproximar mais do valor que a ação poderá assumir. A equação característica de um modelo APT é multifatorial, vários betas e esses betas são encontrados por análise de regressão. A análise de regressão multivariada é utilizada para investigar e modelar a relação estatística entre uma variável resposta, no caso a IBOVESPA e variáveis preditoras (PIB, taxa de juros, entre outras).

1.1 OBJETIVOS

1.1.1 Geral

Aplicar a Teoria de Precificação por Arbitragem, *Arbitrage Pricing Theory* (APT), para estimar um índice de retorno de mercado após uma crise financeira.

1.1.2 Específicos

- Verificar os componentes do modelo *APT* que incluem os fatores influenciados pela crise econômica financeira;
- Quantificar os fatores específicos de cada mercado estudado.

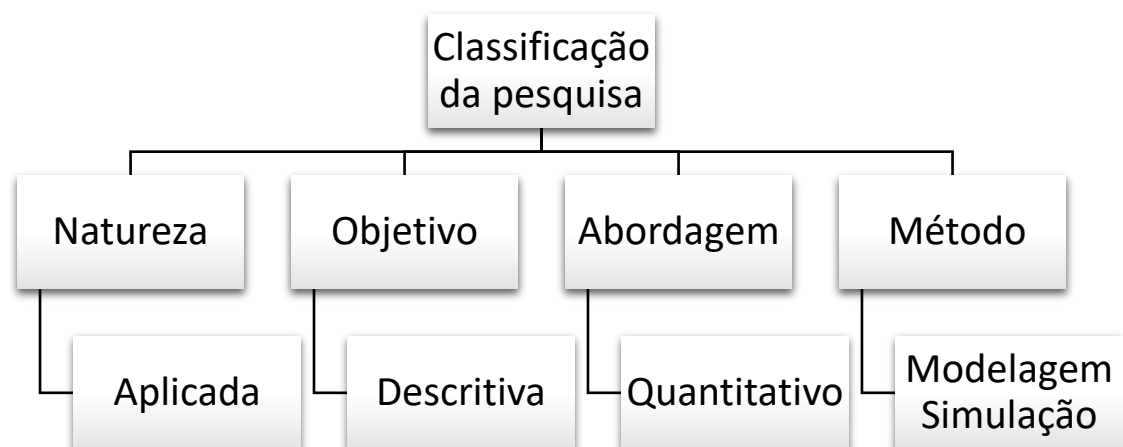
1.2 JUSTIFICATIVA

Com a pesquisa bibliográfica é possível verificar que o modelo *APT* consegue relacionar as variáveis Macroeconômicas e o retorno do ativo ou portfólio, resta agora testar para o caso de mercados em crise como na situação atual do mercado brasileiro. Por causa, justamente, de um mercado ter crises que se espera que o modelo multifatorial, que tenha mais de um beta, seja capaz de se aproximar mais da situação real do que o valor da taxa vai assumir. A ideia não é acertar o valor da taxa em cheio mas conseguir fazer a estimativa dependendo da dinâmica do mercado (crise do mercado).

1.3 MÉTODOS DE PESQUISA

A pesquisa é de natureza aplicada. Pode ser classificada como descritiva quanto ao objetivo e como quantitativa em relação à abordagem. O método adotado foi modelagem e simulação. A figura 1 ilustra esquematicamente a classificação da pesquisa quanto aos diversos fatores.

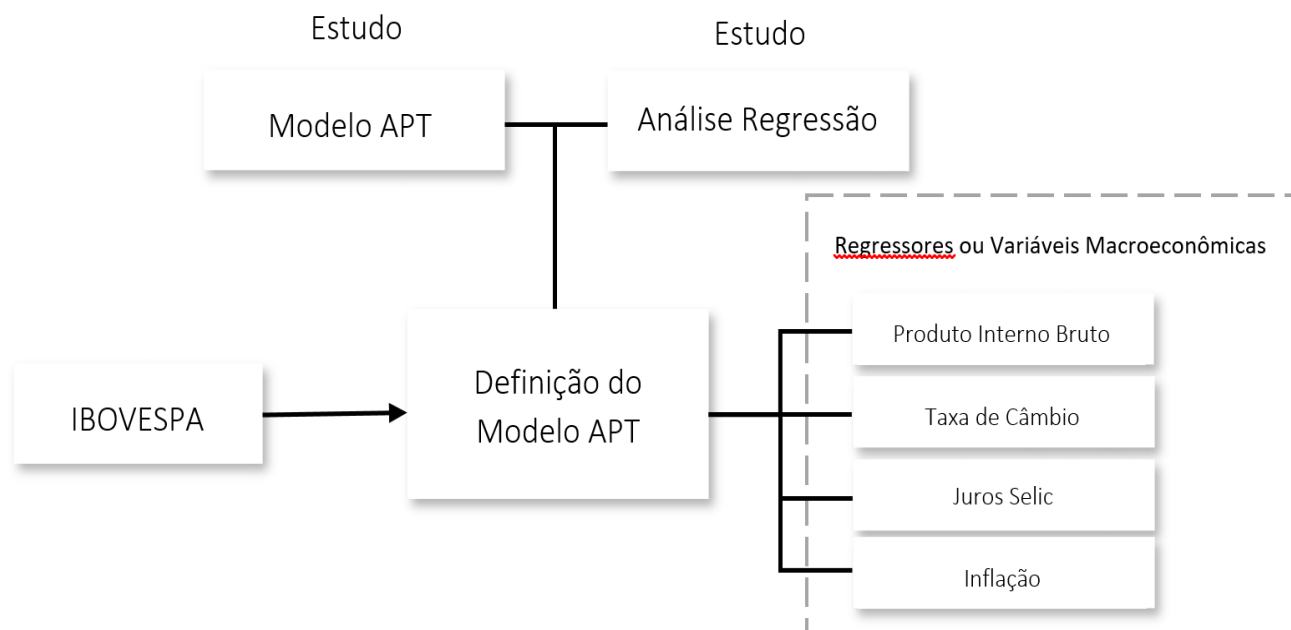
Figura 1 – Classificação da Pesquisa



Fonte: Produção do próprio autor

A abordagem metodológica de pesquisa adotada é apresentado na figura 2. A primeira etapa da pesquisa é a análise bibliométrica sobre Teoria de Precificação por Arbitragem; cálculo do regressor β ; análise fatorial; e definição do Modelo APT. Em seguida, é estabelecida a coleta de dados de sites conhecidos e sobre eles calculado os retornos. Foi feita a análise dos resultados verificando se o modelo APT desenvolvido era capaz de definir o retorno para o Índice da Bolsa de Valores de São Paulo (IBOVESPA) considerando as várias fontes de risco. Por fim, são comparados os resultados e retiradas as conclusões.

Figura 2 – Abordagem metodológica desta Pesquisa



Fonte: Produção do próprio autor

1.4 ESTRUTURA DO TRABALHO

O trabalho está dividido em cinco capítulos. O primeiro capítulo contém a introdução, os objetivos, os métodos da pesquisa e a estrutura do trabalho. A Teoria de Precificação por Arbitragem e sua comparação com o Modelo de Precificação de Ativos, *Capital Asset Pricing Theory* – CAPM, mais a definição dos seus fatores e as medidas de dependência entre suas variáveis são estudadas no segundo capítulo. O terceiro capítulo abrange os modelos de regressão múltipla, com sua análise, suas medidas de desempenho e suas etapas de desenvolvimento de um modelo de previsão. A coleta de dados, a aplicação da regressão múltipla na estimação de um modelo APT e a análise dos resultados são abordadas no quarto capítulo. Por fim, no quinto e último capítulo, as conclusões são feitas, e também há sugestões sobre futuras pesquisas.

2 A TEORIA DE PRECIFICAÇÃO POR ARBITRAGEM – ATP

2.1 CONSIDERAÇÕES INICIAIS

A Teoria de Precificação por Arbitragem, *Arbitrage Pricing Theory* – APT, mede a relação entre o risco de um ativo e o seu retorno exigido. Foi primeiramente publicada por Stephen A. Ross em Dezembro de 1976 – *The arbitrage theory of capital asset pricing*, no *Journal of Economic Theory*. Ela expande e questiona o Modelo de Precificação de Ativos, *Capital Asset Pricing Theory* – CAPM, o qual foi e é utilizado amplamente na busca pela precificação de ativos.

Para Lencione (2005), a discussão sobre precificação de ativos iniciou-se com Markowitz, primeiro autor a analisar o assunto, o qual buscava uma seleção de ativos para um “*portfolio*”. Markowitz (1952) observou que, num plano entre retorno esperado e risco, os ativos se distribuem de tal forma que uns dominam outros, no sentido de que têm maior retorno esperado para um mesmo nível de risco. Assim, as escolhas entre ativos (ou combinações de ativos, a que são chamadas de carteiras) dependeriam dos graus de aversão a risco dos investidores, que passariam a escolher suas carteiras ao longo de uma região chamada “fronteira eficiente”, nos quais se encontram os ativos (ou combinações entre eles) que “dominam” os demais, com os melhores retornos para cada nível de risco (“*portfolios*” eficientes).

A fronteira eficiente era obtida, ao se representar, num gráfico de retorno esperado “versus” desvio-padrão, todas as combinações de ativos e reconhecendo que os investidores são racionais, isto é, sempre preferirão mais riqueza a menos riqueza, para o mesmo nível de risco. Existiam, no entanto, duas dificuldades: primeiro, a quantidade elevada de cálculos necessários para se chegar à fronteira eficiente de Markowitz e em segundo, o fato de que o Modelo Markowitz ser um modelo geral, aplicado à diversificação de ativos de uma carteira e à sua composição, mas não especificamente à precificação de um ativo.

Portanto, se introduziram os conceitos de risco total, risco sistemático e risco não sistemático.

Como Ross (1976) observou, a taxa de retorno de qualquer ação negociada num mercado financeiro é formada por duas partes. A primeira é o retorno normal ou esperado da ação, aquela parte do retorno que os investidores preveem ou esperam. Depende de todas as informações que os investidores possuem a respeito da ação, e usa todo o conhecimento sobre o que poderá influenciar a ação no próximo mês. A segunda parte é o retorno incerto da ação. Esta é a parcela

que decorre de informação que será revelada durante o mês. Pode-se escrever a taxa de retorno no próximo mês de tal modo:

$$R = \bar{R} + U \quad (1)$$

Onde,

R é o retorno total efetivo do mês, $\bar{R}$ é a parte esperada do retorno e U é a parcela inesperada do retorno.

A parcela esperada de qualquer anúncio faz parte da informação que o mercado utiliza para formar a expectativa $\bar{R}$, da taxa de retorno da ação. A surpresa é a novidade que influencia o retorno inesperado da ação U .

Assim, a parcela inesperada do retorno – aquela que resulta de surpresas – é o verdadeiro risco de qualquer investimento. Existem diferenças importantes, no entanto, entre as várias fontes de risco. Algumas fontes de risco dizem respeito especificamente à uma empresa, e outras são mais genéricas, afetando toda a economia. Sendo assim, divide-se esses dois tipos de fontes de riscos em dois componentes: uma parcela sistemática, chamada de risco sistemático, e o restante, que se chama de risco específico ou não sistemático.

Um risco sistemático é qualquer risco que afeta um grande número de ativos, cada um com maior ou menor intensidade. Já um risco não sistemático é um risco que afeta especificamente um único ativo ou um pequeno grupo de ativos.

A incerteza a respeito de condições econômicas gerais, tais como PNB, taxas de juros ou inflação, é um exemplo de risco sistemático. Essas condições afetam praticamente todas as ações de alguma maneira. Essas forças às quais todas as empresas são suscetíveis formam a essência do risco sistemático. Em contrapartida, para se dizer que uma informação é não sistemática e afeta apenas algumas empresas específicas, diz-se às vezes que se trata de um risco idiossincrático.

A diferenciação de um risco sistemático e de um risco não sistemático não é tão exata como parece. Mesmo a notícia mais centrada e peculiar a respeito de uma empresa pode exercer efeitos em toda economia. Porém isto não será levado em consideração.

Voltando a equação (1) pode-se dividir a parcela inesperada U em seus dois componentes: o sistemático e o não sistemático. Como de costume, usa-se a letra grega épsilon, ε , para representar o risco não sistemático e a letra m para indicar o risco sistemático (em algumas situações denominado risco de mercado destacando a sua influência em todos os ativos do mercado com alguma intensidade):

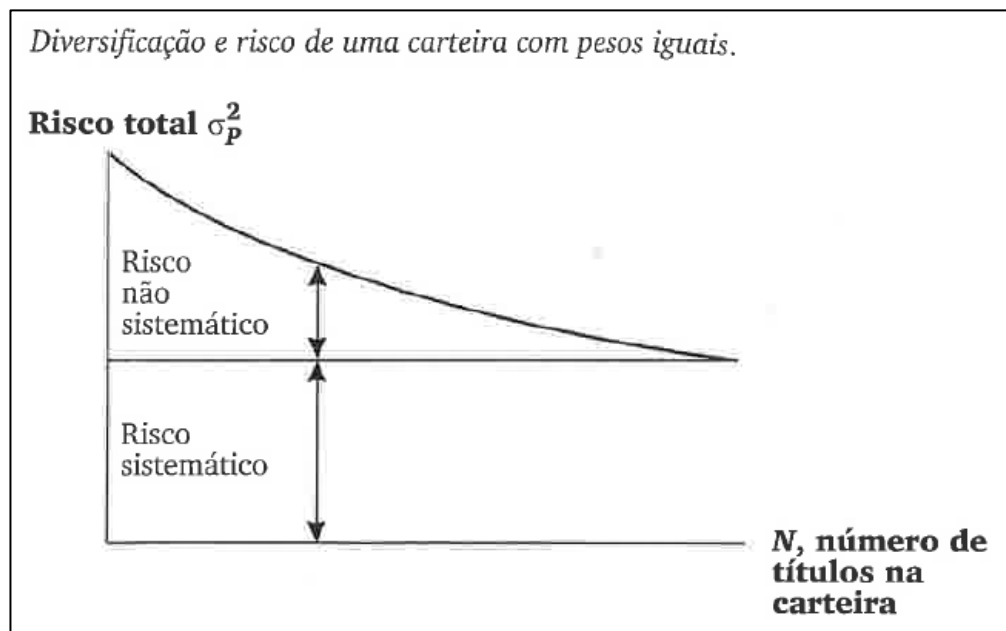
$$R = \bar{R} + U$$

$$R = \bar{R} + m + \varepsilon \quad (2)$$

O fato de que as partes não sistemáticas dos retornos de duas empresas não têm relação uma com a outra não significa que também não haja relação com as partes sistemáticas. Ao contrário, como ambas as empresas são influenciadas pelos mesmos riscos sistemáticos, os riscos sistemáticos de empresas individuais e, portanto, seus retornos totais, tenderão a ser correlacionados.

Em se tratando de uma carteira eficiente, o risco não sistemático tende a sumir, pois esses riscos são independentes um do outro, o efeito da diversificação torna-se mais forte à medida que são acrescentados mais ativos nela. A carteira resultante torna-se menos arriscada, sua taxa de retorno será cada vez mais certa. Porém, o risco sistemático o qual afeta todos os títulos não pode ser evitado em aplicações em muitos títulos, pois não há efeito de diversificação nesse caso. A figura 3 exemplifica a diversificação e o risco numa carteira, podendo ser observado que o risco total cai à medida que aumenta o número de títulos. Esta queda apenas acontece na parcela não sistemático.

Figura 3 – Gráfico Risco total *versus* Número de títulos na carteira



Fonte: Ross, *et al.* (2013)

Por exemplo, uma mudança na taxa de inflação influenciará quase todas as empresas de alguma forma. Se a ação de uma empresa tender a subir quando for anunciado que a inflação

está maior que as expectativas, diz-se que está positivamente associada à inflação. Se cair, no mesmo caso de alta inflacionária, e subir quando a inflação cair, diz-se que está negativamente relacionada à inflação.

A influência de um risco sistemático tal como a inflação sobre uma ação é representada pelo coeficiente beta. O coeficiente beta, β , mede a sensibilidade do retorno da ação a um risco sistemático.

Por exemplo, supôs-se que havia sido identificado três riscos sistemáticos. Estes, no caso, são suficientes para descrever os riscos sistemáticos que influenciam os retornos de ações. Inflação, o PNB (Produto Interno Bruto) e as taxas de juros são bons três candidatos, portanto, cada ação terá um beta associado a cada um desses riscos sistemáticos: um beta da inflação, um beta de PNB e um beta de taxas de juros, juntando as equações (1) e (2):

$$\begin{aligned} R &= \bar{R} + U \\ R &= \bar{R} + m + \varepsilon \\ R &= \bar{R} + \beta_I F_I + \beta_{PNB} F_{PNB} + \beta_r F_r + \varepsilon \end{aligned} \quad (3)$$

Onde, os símbolos β_I representa o beta de inflação da ação, β_{PNB} o beta de Produto Nacional Bruto, e β_r o beta de taxas de juros, F indica uma surpresa, seja de inflação, PNB ou taxas de juros. (Ross, et al. (2013))

O modelo analisado anteriormente é chamado de modelo fatorial, e as fontes de risco sistemático, representadas por F , são os fatores. Formalmente, um modelo de k fatores é um modelo no qual a taxa de retorno de cada ação é obtida por:

$$R = \bar{R} + \beta_1 F_1 + \beta_2 F_2 + \dots + \beta_k F_k + \varepsilon \quad (4)$$

2.2 APT x CAPM

O CAPM, *Capital Asset Pricing Model* (Modelo de Precificação de Ativos) foi desenvolvido no início da década de 1960, simultaneamente, por Sharpe (1964), Litner (1965), e Mossin (1966), é atualmente bastante usado no mercado de capitais, já o APT é mais recente e uma opção aos métodos de precificação de ativos.

Segundo Ross, *et al.* (2013), as diferenças entre os dois modelos são encontradas no modo do tratamento dado pela APT à inter-relação dos retornos dos títulos. Essa não é a única

diferença entra as hipóteses dos dois modelos. Por exemplo, o CAPM geralmente supõe que os retornos dos ativos têm distribuição normal ou que os investidores possuem funções utilidade quadráticas. A APT não exige nenhuma dessas hipóteses.

Ainda segundo Ross, Westerfield, *et al.* (2013), a APT pressupõe que os retornos dos títulos são gerados por uma série de fatores setoriais e gerais do mercado. Há correlação entre os retornos de dois títulos quando são influenciados pelo mesmo fator, ou fatores. Em contrapartida, embora o CAPM permita a existência de correlação entre títulos, não especifica os fatores causadores dessa correlação. Tanto a APT quanto o CAPM preveem uma relação positiva entre retorno esperado e risco. A APT permite que esta relação seja desenvolvida de uma maneira particularmente intuitiva. Além disso, a APT encara o risco de uma maneira mais geral do que simplesmente como covariância padronizada, ou beta de um título em relação à carteira de mercado. No CAPM, o beta de um título mede apenas a sensibilidade do título a flutuações da carteira de mercado. Esse beta determina o retorno de um ativo como sendo a razão da covariância dos retornos do próprio ativo e do retorno de mercado, com o desvio padrão do retorno de mercado.

Nas equações (5) e (6), é ilustrado o cálculo geral e o beta do modelo CAPM:

$$R_i = R_f + \beta_i(R_m - R_f) \quad (5)$$

$$\beta_i = \frac{Cov(R_i, R_m)}{\sigma^2(R_m)} \quad (6)$$

Onde,

β_i é o beta relacionado ao ativo no modelo CAPM

R_f é a taxa livre de risco do ativo

R_i é o retorno do próprio ativo

R_m é o retorno de mercado

A carteira de mercado para que se obtenha o retorno de mercado, R_m , é definida por apenas uma carteira a qual incluiria todos os ativos existentes no mundo. Porém, muitos autores defendem que essa carteira usada para definir R_m , com boa aproximação, possa ser um dos índices de mercado (*proxys*). No Brasil, pode-se citar o IBOVESPA como uma boa *proxy* do mercado brasileiro.

2.3 DEFINIÇÃO DAS VARIÁVEIS OU FATORES EM UM MODELO APT

Os pesquisadores ainda não chegaram num consenso a respeito de qual é o conjunto correto de fatores. O número de k fatores leva em consideração fatores econômicos e financeiros tais como crescimento do PNB – Produto Nacional Bruto, taxa de inflação, força da economia mundial, entre outros.

Segundo Dhyrnes, *et al.* (1984) o número k de fatores dependerá diretamente do tamanho do conjunto de ativos a serem estudados, ou seja, há uma relação de proporcionalidade entre o número de ativos utilizados na pesquisa empírica e o número de fatores a serem estimados.

Já para Rebeschini e Leal (2016), o APT assume que o risco sistemático pode ser medido de diversos modos, mas não especifica exatamente qual ou quantos fatores de risco sistemático são necessários. A escolha dos fatores observados está no critério do analista. Existem três alternativas para a escolha das variáveis a serem usadas como fatores na estimativa do ATP. A primeira são as variáveis microeconômicas, relacionadas a atributos específicos de cada ativo (isto é, dividendos, tamanho da companhia, indústria, etc.). A segunda é extrair os fatores via procedimentos estatísticos, como sendo o principal componente de análises, de suficientes portfólios diversificados para se refletir os riscos sistemáticos do mercado de ações, como indicou Roll e Ross (1980). A terceira é usar variáveis macroeconômicas (inflação, taxas de juros, risco de crédito, etc.) como fatores de risco, como é visto em Chen *et al.* (1986). A vantagem da última aproximação está no fato que ela oferece uma gama de fatores que possibilitam uma interpretação econômica a ser feita, tanto para a exposição ao risco quanto o risco a prêmios a serem pagos.

2.4 MEDIDAS DE DEPENDÊNCIA ENTRE VARIÁVEIS

Rebeschini e Leal (2016) afirmam que o uso de fatores macroeconômicos na estimação do APT deve seguir certas características. O modelo deve considerar os fatores que não sejam correlacionados.

Para Montgomery e Runger (2002), quando duas ou mais variáveis aleatórias são definidas em um espaço probabilístico, é útil descrever como elas variam conjuntamente; isto é, é útil medir a relação entre as variáveis. Uma medida comum de relação entre duas variáveis aleatórias é a covariância. De modo a definir a covariância, necessita-se descrever o valor esperado de uma função de duas variáveis aleatórias $h(X, Y)$. A definição é uma extensão daquela usada para uma função de uma variável aleatória simples.

Valor esperado de uma Função de duas variáveis aleatórias

$$E[h(X, Y)] = \sum \sum h(x, y) f_{XY}(x, y)$$

X, Y discretas

ou

(7)

$$E[h(X, Y)] = \int \int h(x, y) f_{XY}(x, y) dx dy$$

X, Y contínuas

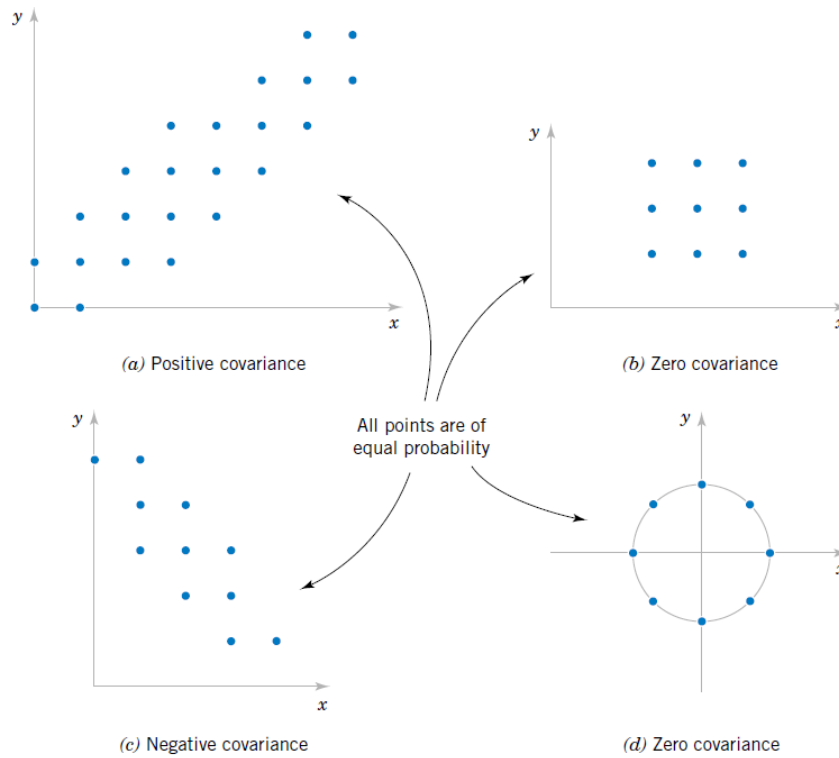
Isto é $E[h(X, Y)]$ pode ser pensado como a média ponderada de $h(x, y)$ para cada ponto na faixa de (X, Y) . O valor de $E[h(X, Y)]$ representa o valor médio de $h(X, Y)$ que é esperado em um longa sequência de tentativas repetidas do experimento aleatório.

A covariância é definida para variáveis contínuas e discretas pela mesma fórmula. A covariância entre as variáveis aleatórias X e Y , denotada por $\text{cov}(X, Y)$ ou σ_{XY} , é

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - \mu_X \mu_Y \quad (8)$$

Se os pontos na distribuição de probabilidades conjuntas de X e Y que recebem probabilidade positiva tenderem a cair ao longo de uma linha de inclinação positiva (ou negativa), então σ_{XY} é positiva (ou negativa). Se os pontos tenderem a cair ao longo de uma linha de inclinação positiva, X tenderá a ser maior que μ_X quando Y for maior que μ_Y . Consequentemente, o produto dos dois termos $x - \mu_X$ e $y - \mu_Y$ tenderá a ser positivo. Entretanto, se os pontos tenderem a cair ao longo de uma linha de inclinação negativa, $x - \mu_X$ tenderá ser positivo, quando $y - \mu_Y$ for negativo e vice-versa. Nesse sentido, a covariância entre X e Y descreve a variação entre duas variáveis aleatórias. A figura 4 mostra exemplos de pares de variáveis aleatórias com covariância positiva, negativa e nula.

Figura 4 – Distribuições de probabilidades conjuntas e o sinal da covariância entre X e Y



Fonte: Montgomery, Runger (2002)

Covariância é uma medida de relação linear entre as variáveis aleatórias. Se a relação entre as variáveis for não-linear, a covariância pode não ser sensível à relação. Isso é ilustrado na figura 4(d). Os únicos pontos com probabilidade não-zero são os pontos no círculo. Há uma relação identificável entre as variáveis, embora a covariância seja zero.

A igualdade das duas expressões para a covariância na equação (8) é mostrada para variáveis aleatórias contínuas, conforme se segue. Escrevendo os valores esperados como integrais

$$\begin{aligned}
 E[(Y - \mu_Y)(X - \mu_X)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x, y) dx dy = \\
 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [xy - \mu_X y - x \mu_Y + \mu_X \mu_Y] f_{XY}(x, y) dx dy
 \end{aligned}$$

Agora

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mu_X y f_{XY}(x, y) dx dy = \mu_X \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy \right] = \mu_X \mu_Y$$

Assim,

$$\begin{aligned}
E[(Y - \mu_Y)(X - \mu_X)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy - \mu_X \mu_Y = E(XY) - \mu_X \mu_Y
\end{aligned}$$

Segundo Montgomery e Runger (2002), há uma outra medida da relação entre duas variáveis aleatórias que é frequentemente mais fácil de interpretar que a covariância.

A correlação entre as variáveis X e Y , denotada por ρ_{XY} , é

$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (9)$$

Devido a $\sigma_X > 0$ e $\sigma_Y > 0$, se a covariância entre X e Y for positiva, negativa ou zero, a correlação entre X e Y será positiva, negativa ou zero, respectivamente. O seguinte resultado pode ser mostrado.

Para quaisquer duas variáveis aleatórias X e Y

$$-1 \leq \rho_{XY} \leq +1 \quad (10)$$

A correlação só escalona a covariância através do desvio-padrão de cada variável. Assim, a correlação é uma grandeza adimensional que pode ser usada para comparar as relações lineares entre pares de variáveis em diferentes unidades.

Se os pontos na distribuição de probabilidades conjuntas de X e Y que recebem probabilidades positivas tenderem a cair ao longo de uma linha de inclinação positiva (ou negativa), então ρ_{XY} será próximo de +1 ou (-1). Se ρ_{XY} for igual a +1 ou -1, pode ser mostrado que os pontos na distribuição de probabilidades conjuntas que recebem probabilidades positivas caem exatamente ao longo de uma linha reta. Duas variáveis aleatórias com correlação não-zero são ditas estar correlacionadas. Similar à covariância, a correlação é uma medida da relação linear entre as variáveis aleatórias.

Para variáveis aleatórias independentes, não esperamos qualquer relação em sua distribuição de probabilidades conjuntas.

Se X e Y forem variáveis aleatórias independentes, então

$$\sigma_{XY} = \rho_{XY} = 0 \quad (11)$$

3 MODELO DE REGRESSÃO MÚLTIPLA

3.1 CONSIDERAÇÕES INICIAIS

Constantemente há a necessidade de se relacionar duas ou mais variáveis quantitativas em estudos de engenharia ou ciências. Assim, a análise de regressão tem sido amplamente utilizada nessas situações retornando bons resultados. Segundo Montgomery e Runger (2002), ela trata-se de uma técnica estatística para modelar e investigar a relação entre variáveis. Em alguns casos há mais de um regressor e estes são chamados de modelos de regressão múltipla.

Todos os modelos e as equações utilizadas foram desenvolvidas por Montgomery e Runger (2002).

3.2 ANÁLISE DE REGRESSÃO

Para Montgomery e Runger (2002), um modelo de regressão que contenha mais de um regressor é chamado de um modelo de regressão múltipla.

De modo geral, a variável dependente ou de resposta, Y , pode estar relacionada a k variáveis independentes ou regressoras. O modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \quad (12)$$

é chamado de modelo de regressão linear múltipla com k variáveis regressoras. Já os parâmetros β_j , $j = 0, 1, \dots, k$, são chamados de coeficientes de regressão. Pode-se constatar que esse modelo descreve um hiperplano no espaço k -dimensional das variáveis regressoras $\{x_j\}$. O parâmetro β_j representa a variação esperada na resposta Y por unidade de variação em x_j , quando todos os outros regressores restantes x_i ($i \neq j$) forem mantidos constantes.

Os modelos de regressão linear múltipla são frequentemente utilizados como funções de aproximações. Pode-se dizer que a verdadeira relação funcional entre Y e $x_1, x_2, \dots, x_k$ não é conhecida, no entanto em certas faixas das variáveis independentes, o modelo de regressão linear é uma aproximação adequada.

Para se estimar os coeficientes de regressão no modelo de regressão múltipla utiliza-se do método dos mínimos quadrados.

Na equação (12), supõe-se que $n > k$ observações estejam disponíveis e seja x_{ij} a i -ésima observação ou nível da variável x_j . As observações são

$$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) \quad i = 1, 2, \dots, n \text{ e } n > k$$

É habitual a tabela 1 a qual apresenta os dados para regressão múltipla.

Tabela 1 – Dados para Regressão Linear Múltipla

Y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	...	x_{nk}

Fonte: Montgomery, Runger (2002)

Cada observação $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$ satisfaz o modelo na equação (12) ou

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \epsilon_i \quad i = 1, 2, \dots, n$$

A função dos mínimos quadrados é

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij})^2 \quad (13)$$

Para minimizar L com relação a $\beta_0, \beta_1, \dots, \beta_k$. As estimativas de mínimos quadrados mínimos de $\beta_0, \beta_1, \dots, \beta_k$ têm de satisfazer

$$\frac{\partial L}{\partial \beta_0} |_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) = 0 \quad (14a)$$

e

$$\frac{\partial L}{\partial \beta_j} |_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij}) x_{ij} = 0 \quad (14b)$$

$$j = 1, 2, \dots, k$$

O estimador de mínimos quadrados $\hat{\beta}$ é a solução para β nas equações

$$\frac{\partial L}{\partial \beta} = \mathbf{0} \quad (19)$$

Deixando de lado os detalhes da obtenção das derivadas e partindo dos seus resultados:

Equações Normais

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \quad (20)$$

As equações (20) são as equações normais de mínimos quadrados na forma matricial. Elas são idênticas à forma escalar das equações normais dadas anteriormente. Com o objetivo de resolver as equações normais, multiplicou-se ambos os lados das Equações (20) pelo inverso de $\mathbf{X}'\mathbf{X}$. Consequentemente, a estimativa de mínimos quadrados de β é

Estimativa de Mínimos quadrados de β

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (21)$$

Observa-se que há $p = k + 1$ equações normais para $p = k + 1$ incógnitas (os valores de $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$). Além disso, a matriz $\mathbf{X}'\mathbf{X}$ é sempre não-singular, como foi considerado anteriormente, de modo que, para inverter essas matrizes, os métodos descritos nos livros-testos sobre determinantes e matrizes podem ser usados para encontrar $(\mathbf{X}'\mathbf{X})^{-1}$. Na prática, os cálculos de regressão múltipla são quase sempre realizados em um computador.

É fácil ver que a forma matricial das equações normais é idêntica à forma escalar. Escrevendo a equação (20) em detalhes

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \dots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \dots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \dots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix}$$

Se a multiplicação matricial indicada for feita, resultará na forma escalar das equações normais. Nessa forma, é fácil ver que $\mathbf{X}'\mathbf{X}$ é uma matriz simétrica ($p \times p$) e $\mathbf{X}'\mathbf{y}$ é um vetor coluna ($p \times 1$). Na matriz $\mathbf{X}'\mathbf{X}$, os elementos da diagonal $\mathbf{X}'\mathbf{X}$ são as somas dos quadrados dos elementos nas colunas de $\mathbf{X}$ e os elementos fora da diagonal são as somas dos produtos cruzados dos elementos nas colunas de $\mathbf{X}$. Além disso, note que os elementos $\mathbf{X}'\mathbf{y}$ são as somas dos produtos cruzados das colunas de $\mathbf{X}$ e das observações de $\{y_i\}$.

O modelo ajustado de regressão é

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \quad i = 1, 2, \dots, n \quad (22)$$

Na notação matricial, o modelo ajustado é

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad (23)$$

A diferença entre a observação y_i e o valor ajustado $\hat{y}_i$ é um resíduo, $e_i = y_i - \hat{y}_i$. O vetor ($n \times 1$) dos resíduos é denotados por

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (24)$$

3.3 MEDIDAS DE DESEMPENHO DE UM MODELO DE REGRESSÃO

Montgomery e Runger (2002) afirmam que modelos de regressão múltipla são quase sempre ajustados em computadores. O computador calcula também outras grandezas que refletem dados importantes acerca do modelo de regressão.

- Estimando σ^2

O estimador σ^2 , a variância do termo do erro ϵ em um modelo de regressão múltipla, é uma dessas grandezas de saída do computador. Esse é um estimador não-tendencioso de σ^2 . Como em regressão linear simples, onde há apenas dois parâmetros no modelo, e em múltipla, onde há p parâmetros, a estimativa de σ^2 é obtida da análise de variância para o modelo de regressão. A equação (25) ilustra esse estimador:

$$\hat{\sigma}^2 = \sum_{i=1}^n e_i^2 / n - p = SQ_E / n - p \quad (25)$$

O numerador da equação (25) é chamado de erro ou de soma dos quadrados dos resíduos e o denominador $n - p$ é chamado de graus de liberdade do erro ou do resíduo.

Pode-se encontrar uma fórmula para calcular SQ_E :

$$SQ_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e} \quad (26)$$

Substituindo $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ na equação (20):

$$SQ_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \quad (27)$$

As propriedades estatísticas dos estimadores de mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ podem ser facilmente encontrados, sujeitas a certas suposições os termos do erro $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ no modelo de regressão. Considera-se que os erros ϵ_i sejam estatisticamente independentes, com média zero e variância σ^2 . Sob essas suposições, os estimadores de mínimos quadrados $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ são estimadores não-tendenciosos dos coeficientes de regressão $\beta_0, \beta_1, \dots, \beta_k$. Essa propriedade pode ser mostrada:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ E(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} \end{aligned} \quad (28)$$

Já que $E(\boldsymbol{\epsilon}) = \mathbf{0}$ e $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$, a matriz identidade. Assim, $\hat{\boldsymbol{\beta}}$ é um estimador não-tendencioso de $\boldsymbol{\beta}$.

As variâncias dos $\hat{\beta}$'s são expressas em termos dos elementos da inversa da matriz $\mathbf{X}'\mathbf{X}$. A inversa de $\mathbf{X}'\mathbf{X}$ vezes a constante σ^2 representa a matriz de covariância dos coeficientes de regressão $\hat{\boldsymbol{\beta}}$. Os elementos da diagonal de $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ são as variâncias de $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ e os elementos fora da diagonal dessa matriz são as covariâncias.

Em geral, a matriz de covariância de $\hat{\boldsymbol{\beta}}$ é uma matriz simétrica ($p \times p$), cujo j -ésimo elemento é a variância de $\hat{\beta}_j$ e cujo i,j -ésimo elemento é a covariância entre $\hat{\beta}_i$ e $\hat{\beta}_j$, ou seja,

$$\text{cov}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C} \quad (29)$$

As estimativas das variâncias desses coeficientes de regressão são obtidas trocando σ^2 por uma estimativa. Quando σ^2 for trocado por sua estimativa $\hat{\sigma}^2$, a raiz quadrada da variância estimada do j -ésimo coeficiente de regressão é chamada de erro-padrão estimado de $\hat{\beta}_j$ ou $ep(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$. Esses erros-padrão são uma medida útil da precisão de estimação para os coeficientes de regressão; erros-padrão pequenos implicam boa precisão.

Programas computacionais de regressão múltipla geralmente apresentam esses erros-padrão. A estimativa da inclinação é cerca de duas vezes a magnitude de seu erro-padrão e $\hat{\beta}_1$ e $\hat{\beta}_2$ são consideravelmente maiores que $ep(\hat{\beta}_1)$ e $ep(\hat{\beta}_2)$. Isso implica precisão razoável de estimação, embora os parâmetros β_1 e β_2 sejam estimados muito mais precisamente do que a interseção (isso é usual em regressão múltipla).

Montgomery e Runger (2002) analisaram que certos testes de hipóteses relativos aos parâmetros do modelo são úteis na medida de adequação de regressão linear múltipla. São eles:

- Teste para a Significância da Regressão

O teste para a significância da regressão é um teste para determinar se há uma relação linear entre a variável de resposta y e um subconjunto de regressores $x_1, x_2, \dots, x_k$. As hipóteses apropriadas são

Hipóteses para o Teste de ANOVA

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \text{ para no mínimo um } j \quad (30)$$

A rejeição de $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ implica que no mínimo uma das variáveis regressoras $x_1, x_2, \dots, x_k$ contribui significativamente para o modelo.

O teste para a significância da regressão é uma generalização do procedimento usado na regressão linear simples. A soma total dos quadrados SQ_T é dividida na soma dos quadrados devido à regressão e na soma dos quadrados devido ao erro.

$$SQ_T = SQ_R + SQ_E \quad (31)$$

No entanto, se $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ for verdadeira, então SQ_R/σ^2 será uma variável aleatória qui-quadrado, com k graus de liberdade. Perceba que o número de graus de liberdade para essa variável qui-quadrado é igual ao número de variáveis regressoras no modelo. Pode-se também mostrar que SQ_E/σ^2 é uma variável aleatória qui-quadrado, com $n - p$ graus de liberdade e que SQ_E e SQ_R são independentes. A estatística de teste para $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ é

Estatística de Teste para ANOVA

$$F_0 = \frac{SQ_R/k}{SQ_E/(n-p)} = \frac{MQ_R}{MQ_E} \quad (32)$$

Deve-se rejeitar H_0 se o valor calculado da estatística de teste na equação (32), f_0 , for maior do que $f_{\alpha, k, n-p}$. O procedimento é geralmente resumido em uma tabela de análise de variância, como a tabela 2.

Tabela 2 – Análise de Variância para Testar a Significância na Regressão Múltipla

Fonte de Variação	Soma dos Quadrados	Graus de Liberdade	Média Quadrática	F_0
Regressão	SQ_R	k	MQ_R	MQ_R / MQ_E
Erro ou resíduo	SQ_E	$n - p$	MQ_E	
Total	SQ_T	$n - 1$		

Fonte: Montgomery, Runger (2002)

Uma fórmula de cálculo para SQ_R pode ser encontrada facilmente. Agora, uma vez que $SQ_T = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n = \mathbf{y}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n$, pode-se escrever as equações (33) e (34)

$$SQ_E = \mathbf{y}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n - \left[\hat{\beta}'\mathbf{X}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n \right] \quad (33)$$

ou

$$SQ_E = SQ_T - SQ_R$$

Desse modo, a soma dos quadrados da regressão é

$$SQ_R = \hat{\beta}'\mathbf{X}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n \quad (34)$$

A maioria dos programas computacionais de regressão múltipla fornece, em sua saída, o teste de significância de regressão, os quais são elucidados por Montgomery e Runger (2009).

- R^2 e R^2 ajustado

Pode-se, também, usar o coeficiente de determinação múltipla R^2 como uma estatística global para avaliar o ajuste do modelo.

Computacionalmente,

$$R^2 = \frac{SQ_R}{SQ_T} = 1 - \frac{SQ_E}{SQ_T} \quad (35)$$

A estatística R^2 é de algum modo problemática como uma medida da qualidade do ajuste para um modelo de regressão múltipla, uma vez que ela sempre aumenta quando a variável é adicionada a um modelo. Em muitos estudos de regressão tem se preferido usar a estatística R^2 ajustado.

R^2 ajustado

$$R^2_{ajustado} = 1 - \frac{SQ_E/(n-p)}{SQ_T/(n-1)} \quad (36)$$

Uma vez que $SQ_E/(n-p)$ é a média quadrática do erro ou do resíduo e $SQ_T/(n-1)$ é uma constante, $R^2_{ajustado}$ somente aumentará quando uma variável for adicionada ao modelo, se a nova variável reduzir a média quadrática do erro.

A estatística R^2 ajustado penaliza essencialmente o analista pela adição de termos ao modelo. É uma maneira fácil de se resguardar contra ajuste em excesso; ou seja, a inclusão de regressores que não são realmente úteis. Consequentemente, é muito útil comparar e avaliar modelos competitivos de regressão.

- Testes para os Coeficientes Individuais de Regressão e Subconjuntos de Coeficientes

Montgomery e Runger (2002), também localizaram em seus estudos que há sempre interesse em testar hipóteses para os coeficientes individuais de regressão. Tais testes são úteis na determinação do valor potencial de cada uma das variáveis regressoras no modelo de regressão. Por exemplo, o modelo pode ser mais efetivo com a inclusão de variáveis adicionais ou talvez com a retirada de um ou mais regressores atualmente no modelo.

A hipótese para testar se um coeficiente individual de regressão, como β_j , é igual a um valor β_{j0} é

$$\begin{aligned} H_0: \beta_j &= \beta_{j0} \\ H_1: \beta_j &\neq \beta_{j0} \end{aligned} \quad (37)$$

A estatística de teste para essa hipótese é

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{ep(\beta_{j0})} \quad (38)$$

Em que C_{jj} é o elemento da diagonal de $(\mathbf{X}'\mathbf{X})^{-1}$ correspondente a $\hat{\beta}_j$. Nota-se que o denominador da equação (35) é o erro-padrão do coeficiente $\hat{\beta}_j$. A hipótese nula $H_0: \beta_j = \beta_{j0}$ será rejeitada se $|t_0| > t_{\alpha/2, n-p}$. Isso é chamado de teste parcial ou marginal, porque o coeficiente de regressão $\hat{\beta}_j$ depende de todos os outros regressores x_i ($i \neq j$) que estão no modelo.

Para Montgomery e Runger (2002), um caso especial importante da hipótese prévia ocorre para $\beta_{j0} = 0$. Se $H_0: \beta_j = 0$ não for rejeitada, isso indicará que o regressor x_i pode ser retirado do modelo. A adição de uma variável a um modelo de regressão sempre aumenta a soma dos quadrados da regressão e diminui a soma dos quadrados do erro (essa é a razão pela qual R^2 sempre aumenta quando uma variável é adicionada). Deve-se decidir se o aumento da soma dos quadrados da regressão é grande o suficiente para justificar o uso da variável adicionada no modelo. Além disso, a adição de uma variável não importante ao modelo pode realmente aumentar a média quadrática do erro, indicando que a adição de tal variável tornou realmente mais pobre o ajuste do modelo aos dados (essa é a razão pela qual R^2_{ajustado} é uma medida melhor do ajuste global do modelo do que R^2 comum).

Ainda sobre Montgomery e Runger (2002), existe uma outra maneira de testar a contribuição de uma variável regressora individual para o modelo. Essa abordagem determina o aumento na soma dos quadrados da regressão, obtido pela adição de uma variável x_i ao modelo, dado que outras variáveis x_i ($i \neq j$) já estão incluídas na equação de regressão.

O procedimento usado para fazer isso é chamado de teste geral de significância da regressão ou método da soma extra dos quadrados. Esse procedimento pode ser usado também para investigar a contribuição de um subconjunto de variáveis regressoras ao modelo. Considerando o modelo de regressão com k variáveis regressoras

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (39)$$

sendo $\mathbf{y}$ ($n \times 1$), $\mathbf{X}$ ($n \times p$), $\boldsymbol{\beta}$ ($p \times 1$), $\boldsymbol{\epsilon}$ ($n \times 1$) e $p = k + 1$. Deve-se determinar se o subconjunto de variáveis regressoras $x_1, x_2, \dots, x_r$ ($r < k$), como um todo, contribui significativamente para o modelo de regressão. Divide-se o vetor dos coeficientes de regressão conforme se segue:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \quad (40)$$

Em que $\boldsymbol{\beta}_1$ é ($r \times 1$) e $\boldsymbol{\beta}_2$ é $[(p - r) \times 1]$. Deseja-se testar as hipóteses:

Hipóteses para o Teste Geral de Regressão

$$H_0: \boldsymbol{\beta}_1 = \mathbf{0}$$

$$H_1: \boldsymbol{\beta}_1 \neq \mathbf{0} \quad (41)$$

Sendo $\mathbf{0}$ um vetor de zeros. O modelo pode ser escrito como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \quad (42)$$

em que $\mathbf{X}_1$ representa as colunas de $\mathbf{X}$ associadas com $\boldsymbol{\beta}_1$ e $\mathbf{X}_2$ representa as colunas de $\mathbf{X}$ associadas com $\boldsymbol{\beta}_2$.

Para o modelo completo (incluindo de $\boldsymbol{\beta}_1$ como $\boldsymbol{\beta}_2$), sabe-se que $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Em adição, a soma dos quadrados da regressão para todas as variáveis, incluindo a interseção, é

$$SQ_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \quad (p = k + 1 \text{ graus de liberdade})$$

e

$$MQ_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n - p}$$

$SQ_R(\boldsymbol{\beta})$ é chamada de soma dos quadrados da regressão devido a $\boldsymbol{\beta}$. Para encontrar a contribuição dos termos em $\boldsymbol{\beta}_1$ para a regressão, ajusta-se o modelo considerando a hipótese $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ como verdadeira. O modelo reduzido é encontrado da equação (39) como

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \quad (43)$$

A estimativa de mínimos quadrados de $\boldsymbol{\beta}_2$ é $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$ e

$$SQ_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} \quad (p - r \text{ graus de liberdade}) \quad (44)$$

A soma dos quadrados da regressão devido a $\boldsymbol{\beta}_1$, dado que $\boldsymbol{\beta}_2$ já está no modelo, é

$$SQ_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SQ_R(\boldsymbol{\beta}) - SQ_R(\boldsymbol{\beta}_2) \quad (45)$$

Essa soma dos quadrados tem r graus de liberdade. Algumas vezes é chamada de soma extra dos quadrados devido a $\boldsymbol{\beta}_1$. Note que $SQ_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ é o aumento na soma dos quadrados da regressão devido à inclusão das variáveis $x_1, x_2, \dots, x_r$ no modelo. Agora, $SQ_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ é independente de MQ_E , podendo a hipótese nula $\boldsymbol{\beta}_1 = \mathbf{0}$ ser testada pela estatística

Estatística F para o Teste Geral de Regressão

$$F_0 = \frac{SQ_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{MQ_E} \quad (46)$$

Se o valor calculado da estatística de teste for $f_0 > f_{\alpha, r, n-p}$, rejeita-se H_0 , concluindo que no mínimo um dos parâmetros em $\boldsymbol{\beta}_1$ não é zero e, conseqüentemente, no mínimo uma das variáveis $x_1, x_2, \dots, x_r$ em $\mathbf{X}_1$ contribui significativamente para o modelo de regressão. Alguns autores chamam o teste na equação (46) de um teste F parcial.

O teste F parcial é muito útil. Pode-se usá-lo para medir a contribuição de cada regressor individual x_j , como se ele fosse a última variável adicionada ao modelo, calculando $SQ_R(\beta_j|\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$, $j = 1, 2, \dots, k$

Esse é o aumento na soma dos quadrados da regressão devido à adição de x_j ao modelo que já inclui $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k$. O teste F parcial é um procedimento mais geral em que se pode medir o efeito de conjuntos de variáveis.

3.4 ETAPAS NO DESENVOLVIMENTO DE UM MODELO DE PREVISÃO APLICANDO REGRESSÃO MÚLTIPLA

Em modelos de regressão múltipla, para Montgomery e Runger (2002), é frequentemente útil construir estimativas de intervalos de confiança para os coeficientes de regressão $\{\beta_j\}$. O desenvolvimento de um procedimento para obter esses intervalos de confiança requer que os erros $\{\epsilon_i\}$ sejam normal e independentemente distribuídos, com média zero e variância σ^2 . Essa é a mesma suposição requerida no teste de hipóteses. Logo, as observações $\{Y_i\}$ são normal e independentemente distribuídas. Com média $\beta_0 + \sum_{j=1}^k \beta_j x_{ij}$ e variância σ^2 . Uma vez que o estimador de mínimos quadrados $\hat{\beta}$ é normalmente distribuído com vetor médio β e matriz de covariância $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Então, cada uma das estatísticas

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad j = 0, 1, \dots, k \quad (47)$$

tem uma distribuição t , com $n - p$ graus de liberdade, sendo C_{jj} o jj -ésimo elemento da matriz $(\mathbf{X}'\mathbf{X})^{-1}$ e $\hat{\sigma}^2$ a estimativa da variância do erro, obtida pela equação (25). Isso conduz ao intervalo de confiança $100(1 - \alpha)\%$ para o coeficiente de regressão $\beta_j, j = 0, 1, \dots, k$.

Um intervalo de confiança $100(1 - \alpha)\%$ para o coeficiente de regressão $\beta_j, j = 0, 1, \dots, k$ no modelo de regressão linear múltipla é dado por

$$\hat{\beta}_j - t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \quad (48)$$

Pelo fato de $\sqrt{\hat{\sigma}^2 C_{jj}}$ ser o erro-padrão do coeficiente de regressão $\hat{\beta}_j$, escreve-se também a fórmula de IC como $\hat{\beta}_j - t_{\frac{\alpha}{2}, n-p} ep(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\frac{\alpha}{2}, n-p} ep(\hat{\beta}_j)$.

Também é obtido o intervalo de confiança para a resposta média em um determinado ponto, como $x_{01}, x_{02}, \dots, x_{0k}$. Para estimar a resposta média nesse ponto, define-se o vetor

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

A resposta média nesse ponto é $E(Y|x_0) = \mu_{Y|x_0} = \mathbf{x}'_0\boldsymbol{\beta}$, que é estimada por $\hat{\mu}_{Y|x_0} = \mathbf{x}'_0\hat{\boldsymbol{\beta}}$. Esse estimador é não-tendencioso, uma vez que $E(\mathbf{x}'_0|\hat{\boldsymbol{\beta}}) = \mathbf{x}'_0\boldsymbol{\beta} = E(Y|x_0) = \mu_{Y|x_0}$ e a variância de $\hat{\mu}_{Y|x_0}$ é $V(\hat{\mu}_{Y|x_0}) = \sigma^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0$.

Um IC de $100(1 - \alpha)\%$ para $\hat{\mu}_{Y|x_0}$ pode ser construído a partir da estatística

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\sigma^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \quad (49)$$

Para um modelo de regressão linear múltipla, um intervalo de confiança $100(1 - \alpha)\%$ para a resposta média no ponto $x_{01}, x_{02}, \dots, x_{0k}$ é

$$\begin{aligned} & \hat{\mu}_{Y|x_0} - t_{\frac{\alpha}{2}, n-p} \sqrt{\sigma^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \\ \leq \mu_{Y|x_0} \leq & \hat{\mu}_{Y|x_0} - t_{\frac{\alpha}{2}, n-p} \sqrt{\sigma^2\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \end{aligned} \quad (50)$$

Segundo Montgomery e Runger (2002), um modelo de regressão pode ser usado para prever novas ou futuras observações para a variável de resposta Y , correspondendo a valores particulares das variáveis independentes como $x_{01}, x_{02}, \dots, x_{0k}$. Se $\mathbf{x}'_0 = [1, x_{01}, x_{02}, \dots, x_{0k}]$, então uma estimativa da futura observação Y_0 nos pontos $x_{01}, x_{02}, \dots, x_{0k}$ é

$$\hat{y}_0 = \mathbf{x}'_0\hat{\boldsymbol{\beta}} \quad (51)$$

Intervalo de Previsão

Um intervalo de previsão de $100(1 - \alpha)\%$ para essa futura observação é

$$\begin{aligned} & \hat{y}_0 - t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \\ \leq Y_0 \leq & \hat{y}_0 + t_{\frac{\alpha}{2}, n-p} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \end{aligned} \quad (52)$$

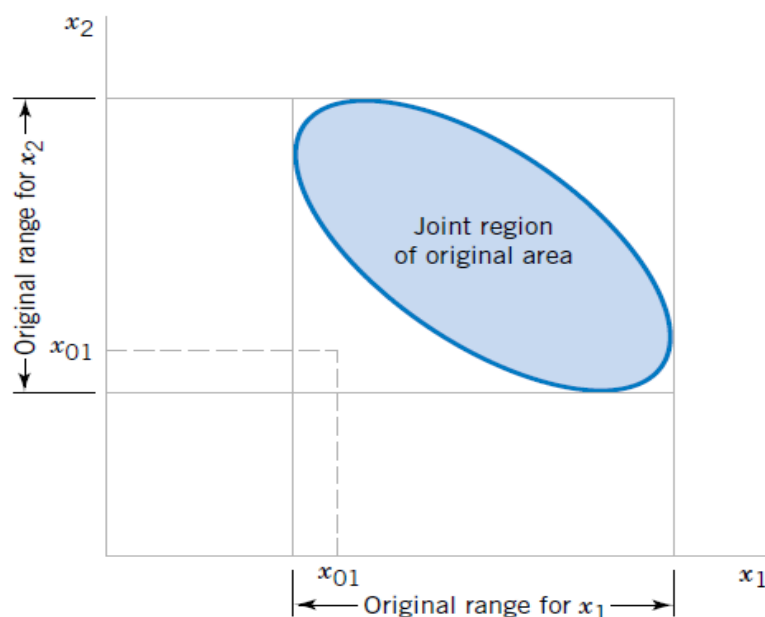
Ao se comparar a equação (52) para o intervalo de previsão com a expressão para o intervalo de confiança para a média, equação (50), irá se observar que o intervalo de previsão é sempre mais largo do que o intervalo de confiança. O intervalo de confiança expressa o erro na estimação da média de uma distribuição, enquanto o intervalo de previsão expressa o erro na previsão de uma observação futura da distribuição no ponto $\mathbf{x}_0$. Ele tem de incluir o erro na

estimação da média naquele ponto, assim como a variabilidade inerente na variável aleatória Y no mesmo valor $\mathbf{x} = \mathbf{x}_0$.

Além disso, alguém pode querer prever a média de vários valores de Y , isto é m , todos no mesmo valor de $\mathbf{x} = \mathbf{x}_0$. Uma vez que a variância de uma amostra é σ^2/m , a equação (52) é modificada como segue. Troca-se a constante 1, sob o sinal de raiz quadrada, por $1/m$ para refletir a menor variabilidade na média de m observações. Isso resulta em um intervalo mais estreito.

Na previsão de novas observações e na estimação da resposta média em um dado ponto $x_{01}, x_{02}, \dots, x_{0k}$, há a necessidade de ter cuidado na extrapolação além da região contendo as observações originais. É bem possível que um modelo que se ajuste bem na região dos dados originais, não se ajuste de forma satisfatória fora daquela faixa. Em regressão múltipla, é frequentemente fácil extrapolar inadvertidamente, uma vez que os níveis das variáveis ($x_{i1}, x_{i2}, \dots, x_{ik}$), $i = 1, 2, \dots, n$, definem conjuntamente a região contendo os dados. Como um exemplo, a figura 4 ilustra a região contendo as observações para um modelo de regressão com duas variáveis. Nota-se que o ponto (x_{01}, x_{02}) está dentro das faixas de ambas as variáveis regressoras x_1 e x_2 , porém está fora da região que é realmente englobada pelas observações originais. Ela é chamada algumas vezes de uma extrapolação disfarçada. Assim, a previsão do valor de uma nova observação ou a estimação da resposta média nesse ponto é uma extrapolação do modelo original de regressão.

Figura 5 – Um exemplo de extrapolação em regressão múltipla



Fonte: Montgomery, Runger (2002)

4 APLICAÇÃO DA REGRESSÃO MÚLTIPLA AO APT

4.1 CONSIDERAÇÕES INICIAIS

Neste capítulo é realizada a aplicação da regressão múltipla na estimação de um modelo APT, levando em conta os fatores escolhidos. Ao final, é possível analisar o modelo pelas medidas de desempenho da regressão e descobrir realmente quais fatores influenciam.

4.2 ESPECIFICAÇÃO DOS FATORES

Após a pesquisa bibliométrica, se obteve diversos artigos científicos em relação ao APT. Dentre os artigos selecionados, foram observadas as variáveis macroeconômicas utilizadas por cada autor, conforme está apresentado na tabela 3.

Tabela 3 – Relação entre os Artigos selecionados e os fatores macroeconômicos abordados em cada um

Artigos	Fatores Macroeconômicos				
	Retorno de Mercado	Taxa de Câmbio	Taxa de Juros	Produto Interno Bruto	Inflação
Ross <i>et al.</i> (1980)	x		x	x	
Basu <i>et al.</i> (2012)		x			
Bilson <i>et al.</i> (2001)		x		x	x
Yli-Olli <i>et al.</i> (1992)			x	x	x
Garcia <i>et al.</i> (2001)	x				x
Rebeschini <i>et al.</i> (2016)			x		x

Fonte: Produção do próprio autor

A partir dos fatores macroeconômicos mais abordados nesses artigos deu-se a escolha dos mesmos para a realização dessa análise. Portanto, os fatores utilizados serão: Inflação, Produto Interno Bruto, Taxa de Juros e Taxa de Câmbio.

No Brasil, a inflação é dada pelo IPCA, Índice de Preços ao Consumidor, é medido mês a mês pelo IBGE, é um índice criado para medir a variação de preços do mercado para o

consumidor final e é como um reflexo do custo de vida de famílias que possuem renda entre um e quarenta salários mínimos, com base em nove regiões metropolitanas do país.

O produto interno bruto (PIB) representa a soma (em valores monetários) de todos os bens e serviços finais produzidos numa determinada região (quer sejam países, estados ou cidades), durante um período determinado (mês, trimestre, ano etc). O PIB é um dos indicadores mais utilizados na macroeconomia com o objetivo de quantificar a atividade econômica de uma região. Analistas indicam a utilização do PIB real para avaliações mais consistentes da variação do PIB, uma vez que este leva em conta apenas as variações nas quantidades produzidas dos bens, e não nas alterações de seus preços de mercado. A análise da variação do PIB exige o uso de um deflator, ou seja, um índice de preços utilizado para descontar o aumento dos preços, isolando o crescimento real das riquezas produzidas.

A Selic é a taxa básica de juros da economia no Brasil, utilizada no mercado interbancário para financiamento de operações com duração diária, lastreadas em títulos públicos federais. A sigla SELIC é a abreviação de Sistema Especial de Liquidação e Custódia. Essa sigla nada mais é que um sistema computadorizado utilizado pelo governo, a cargo do Banco Central do Brasil, para que haja controle na emissão, compra e venda de títulos. A Taxa Selic é obtida pelo cálculo da taxa média ponderada dos juros praticados pelas instituições financeiras.

A taxa de câmbio é o preço de uma moeda estrangeira medido em unidades ou frações (centavos) da moeda nacional. No Brasil, a moeda estrangeira mais negociada é o dólar dos Estados Unidos, fazendo com que a cotação comumente utilizada seja a dessa moeda.

4.3 COLETA DE DADOS

Os fatores utilizados nessa pesquisa foram os índices econômicos: Inflação, Produto Interno Bruto, Taxa de Juros, a Taxa de Câmbio foi deixada de lado na primeira coleta para ser usada caso os indicadores não apresentassem eficiência no modelo. A cada um dos índices deu-se um tratamento anterior.

O conjunto de dados dos fatores, tirando a Taxa de Câmbio, foram coletados no período do dia 25 de Março de 2003 a 31 de Dezembro de 2017, resultando em 3622 dados.

O índice IBOVESPA, utilizado como variável dependente, foi obtida no site *YAHOO! finanças*. Nesse caso, o foco era o rendimento dado por esse índice, assim foi utilizado a equação (52):

$$\text{Rendimento IBOVESPA} = \frac{\text{Valor do fechamento do dia}}{\text{Valor do fechamento do dia anterior}} - 1 \quad (52)$$

A taxa Selic, utilizada como variável independente, foi obtida no site do Banco Central do Brasil, como seus valores são dados anuais, foi utilizada a equação (53) de juros compostos, levando em consideração o ano de 252 dias úteis.

$$\text{Valor futuro} = \text{Valor presente}(1 + \text{taxa de juros})^{\text{número de períodos}} \quad (53)$$

Os dados do PIB real, utilizado como variável independente, foram obtidos no site do IBGE – Instituto Brasileiro de Geografia e Estatística, como seus valores são dados anuais, foi utilizada a sua média aritmética, levando em consideração o ano de 252 dias úteis.

$$\text{Valor diário do PIB} = \frac{\text{Valor anual do PIB}}{252} \quad (54)$$

A dados do IPCA, utilizada como variável independente, foram obtidos no site do IBGE – Instituto Brasileiro de Geografia e Estatística, como seus valores são dados mensais, foi utilizada a sua média aritmética, levando em consideração o mês de 30 dias úteis.

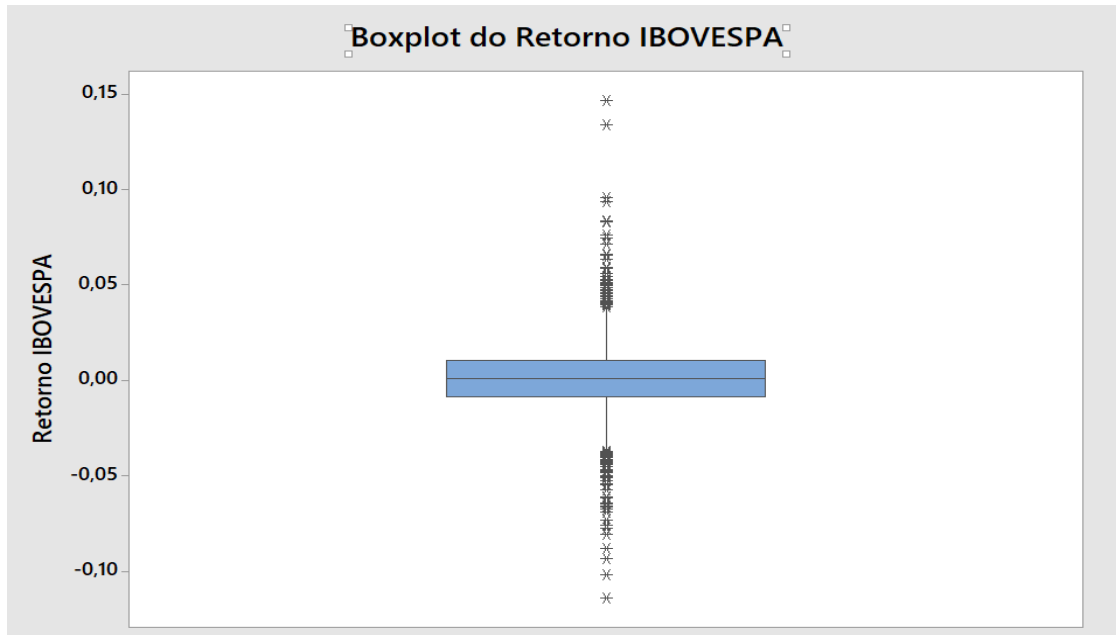
$$\text{Valor diário do IPCA} = \frac{\text{Valor mensal do IPCA}}{30} \quad (54)$$

Todos os valores coletados foram tabelados e aplicou-se logaritmo natural para se obter uma capitalização contínua, resultando em melhores resultados.

4.4 APLICAÇÃO DA REGRESSÃO MÚLTIPLA NA ESTIMAÇÃO DE UM MODELO APT

Antes de aplicar a regressão múltipla foi necessário uma pré análise dos dados da Inflação, do Produto Interno Bruto e da Taxa de Juros. Essa pré análise resultou num gráfico *boxplot* do Retorno IBOVESPA em relação aos demais fatores, o qual ilustrou a precisão modelo e mostrou haver muitos *outliers* (figura 6). Sendo assim, os pontos *outliers* foram retirados para evitar o aumento da dispersão do modelo, tornando-o um modelo melhor e com mais precisão.

Figura 6 – Boxplot do Retorno IBOVESPA



Fonte: Produção do próprio autor

Essa retirada de *outliers* foi orientada pelos valores os quais o programa Minitab identificava como Residual ou Impróprio. Foram feitas 5 rodadas de retirada dos pontos na primeira etapa. No entanto, os valores de R^2 , R^2 ajustado e R^2 previsão foram insatisfatórios, cerca de 1%.

Assim, houve a necessidade de inserir os dados da Taxa de Câmbio na segunda etapa. O conjunto desses dados foi coletado no período do dia 16 de Outubro de 2007 a 29 de Dezembro de 2017, resultando em 2518 dados.

A Taxa de Câmbio, utilizado como variável independente, foi obtida no site *UOL*. Nesse caso, o foco era o rendimento dado por esse índice, assim foi utilizado a equação (55):

$$\text{Câmbio} = \frac{\text{Valor do fechamento do dia}}{\text{Valor do fechamento do dia anterior}} - 1 \quad (55)$$

Novamente houve a necessidade de se retirar *outliers*. Essa retirada, também, foi orientada pelos valores os quais o programa Minitab identificava como Residual ou Impróprio, assim, foram feitas mais 5 rodadas de retirada dos pontos. No entanto, os valores de R^2 , R^2 ajustado e R^2 previsão continuaram insatisfatórios, cerca de 23%.

Nesse momento, se analisou os resultados de *p-value* e se identificou que as variáveis que apresentavam valores de *p-value* maiores ou iguais a 0,05 não deveriam entrar no modelo, pois não teriam influência no mesmo. Quanto menor o valor de *p-value*, mais influência ela

teria no modelo, portanto, foram retiradas as variáveis SELIC e PIB real e uma nova análise, terceira etapa, com apenas o IPCA diário e o Câmbio foi executada no Minitab.

Novamente houve a necessidade de se retirar *outliers*. Essa retirada, também, foi orientada pelos valores os quais o programa Minitab identificava como Residual ou Impróprio, assim, foi feita 1 rodada de retirada dos pontos e os valores de R^2 , R^2 ajustado e R^2 previsão continuaram insatisfatórios, cerca de 23%.

Todos os resultados de R^2 , R^2 ajustado e R^2 previsão colhidos nas etapas de análises feitas no Minitab estão na tabela 4.

Tabela 4 – Resultados colhidos em cada etapa de análise Minitab

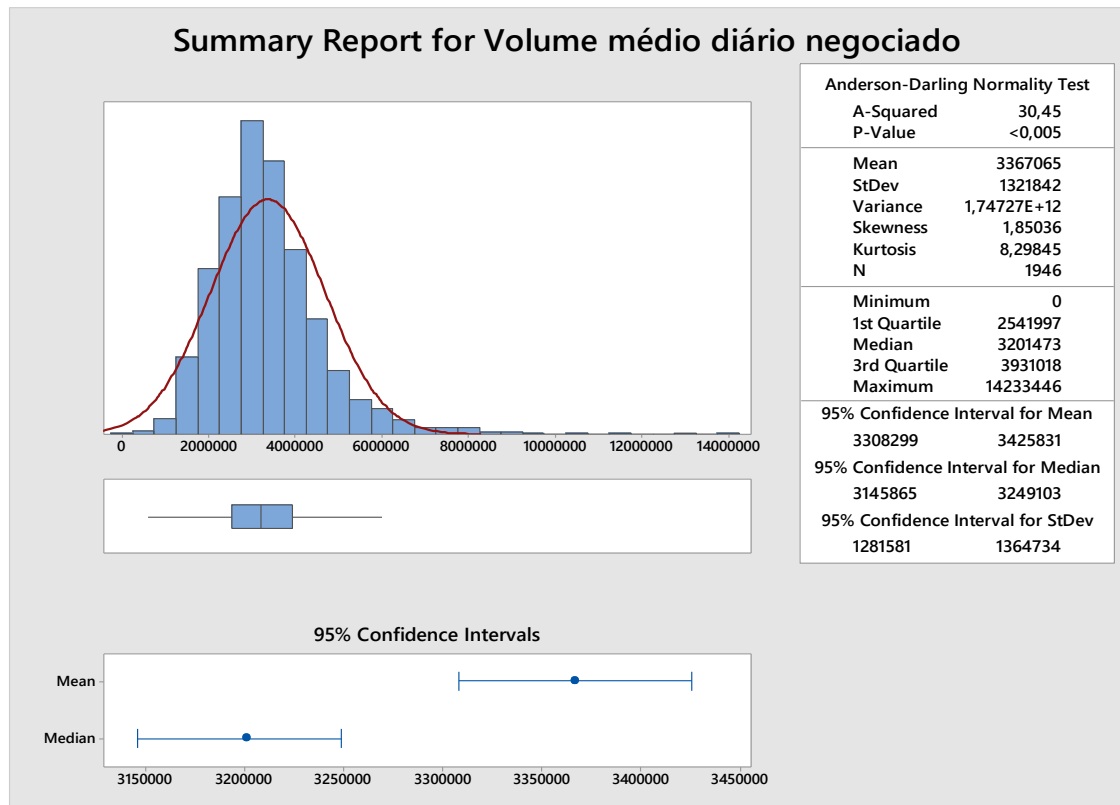
Etapa	Resultados		
	R^2 (%)	R^2 ajustado (%)	R^2 previsão (%)
1	1,11	0,99	0,78
2	23,47	23,33	23,10
3	23,38	23,30	23,15

Fonte: Produção do próprio autor

4.5 ANÁLISE DOS RESULTADOS

Após várias tentativas, as quais citadas acima, houve a necessidade de se incluir mais uma variável no modelo, para se obter um melhor valor de R^2 , R^2 ajustado e R^2 previsão.

Assim, além do IPCA diário e o Câmbio, o modelo passou a contar com a variável *Volume médio diário negociado no Bovespa*, uma variável que expressa a média do volume de negociações no dia, dado em reais, realizados na Bovespa. A figura 7 apresenta a análise da variável *Volume médio diário negociado na Bovespa*.

Figura 7 – Análise da variável *Volume médio diário negociado no Bovespa*

Fonte: Produção do próprio autor

Deste modo, se pode perceber que a variável *Volume médio diário negociado no Bovespa* não pode ser representada como uma distribuição normal, apresenta uma assimetria e tendência a ter uma profundidade maior na cauda da direita, ou seja, período de grandes movimentações. Cabe ressaltar que a movimentação nunca pode ser zero, pois o máximo que pode acontecer é não ter pregão naquele dia, como por exemplo em um feriado nacional.

Os resultados do modelo final estão nas tabelas 5:

Tabela 5 – Resultados de R^2 colhidos no modelo final de análise Minitab

R^2 (%)	R^2 ajustado (%)	R^2 previsão (%)
78,75	77,63	78,42

Fonte: Produção do próprio autor

Os valores de valor de R^2 , R^2 ajustado e R^2 previsão são bem maiores aos encontrados anteriormente, apresentando assim, um melhor ajuste do modelo e a inclusão de regressores que são realmente úteis a este. O valor de R^2 previsão está próximo de R^2 e R^2 ajustado, portanto

o modelo demonstra se ajustar muito bem ao conjunto de dados observados e possui capacidade preditiva adequada. A Tabela 6 apresenta os parâmetros do modelo de regressão.

Tabela 6 – Resultados colhidos no modelo final de análise Minitab

Termo	Coefficiente	<i>p-value</i>	VIF
Constante	0,00738	0,381	
Ln(IPCA)	-8,00	0,003	1,00
Ln(Cambio)	-0,2680	0,000	1,00
Ln(<i>Volume médio diário negociado no Bovespa</i>)	-0,000363	0,020	1,00

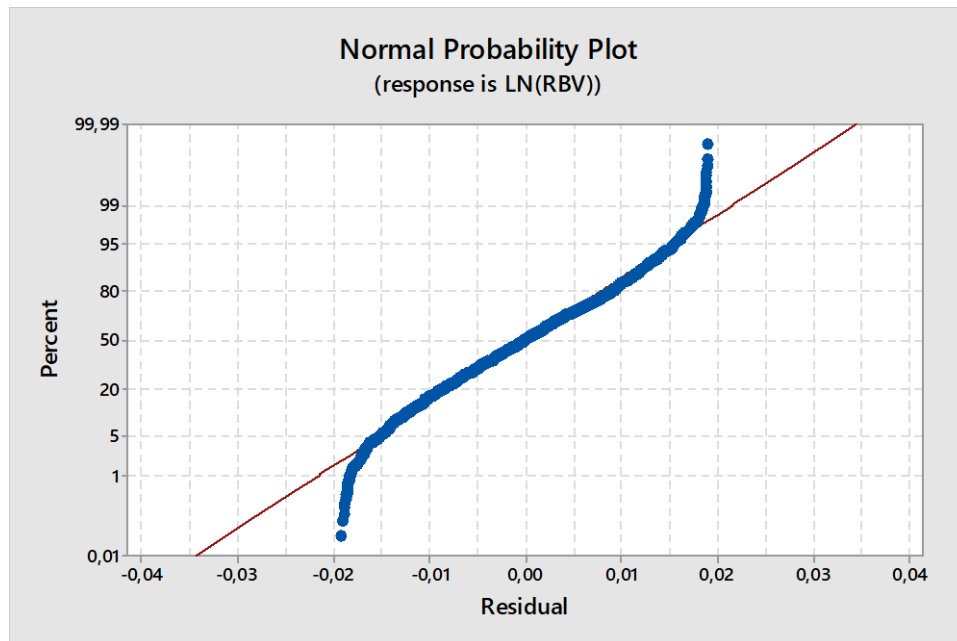
Fonte: Produção do próprio autor

Os valores de *p-value* para os coeficientes estimados, conforme apresentado na tabela 6, mostram que o modelo estimado pelo processo da regressão é significativo em um α -level de 0,05, indicando que pelo menos um coeficiente é diferente de zero.

Conclui-se da análise da tabela acima que os fatores IPCA diário, Câmbio e variável *Volume médio diário negociado no Bovespa* são importantes ao modelo, pois possuem *p-values* menores que 0,05, como era esperado. Se fossem maiores não seriam tão importantes e poderiam ser descartados.

Os valores de VIF (*Variance inflation factors*), fatores de inflação de variância, representam a análise de variância e indicam em que medida a multicolinearidade (correlação entre preditores) está presente na análise de regressão. Se o valor de VIF for maior a 1 ele diz que a variável só aumenta a dispersão dos dados, infla a variância. Como os valores obtidos são iguais a 1 isso significa que eles não estão correlacionados, cada variável não está aumentando a dispersão e contribui realmente para o modelo. O gráfico 1 apresenta a tendência dos resíduos serem normalmente distribuídos.

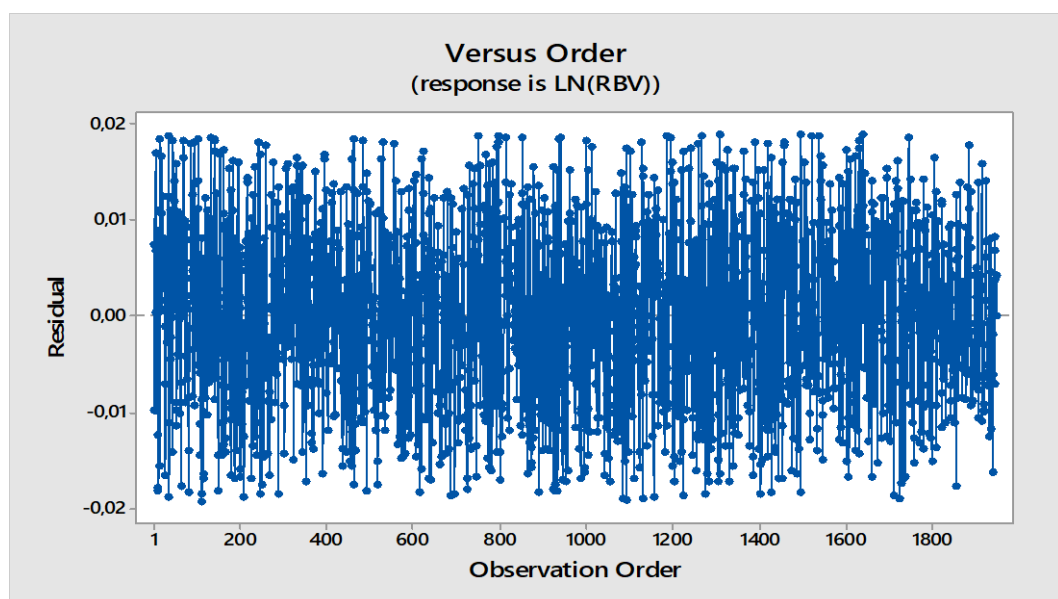
Gráfico 1 – Probabilidade Normal



Fonte: Produção do próprio autor

O gráfico de probabilidade normal mostra um padrão de distribuição dos resíduos aproximadamente linear coerente com uma distribuição normal. Os valores extremos (as caudas no gráfico) não seguem muito bem a distribuição normal, portanto existem outliers que ainda não foram retirados, mas não prejudicam o modelo. O gráfico 2 exibe a dispersão dos resíduos.

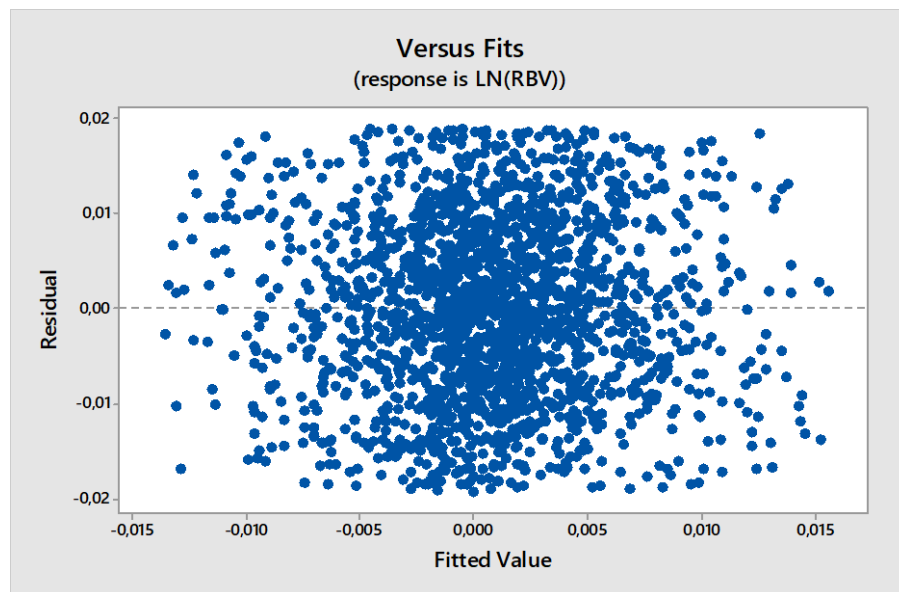
Gráfico 2 – Resíduos versus Ordem



Fonte: Produção do próprio autor

O gráfico de *resíduos versus ordem* mostra os resíduos na ordem em que os dados foram coletados, a sua dispersão. Nesse caso, pode-se afirmar que os resíduos são independentes um do outro porque não mostram tendências nem padrões quando exibidos em ordem temporal. Os resíduos ficam com média 0 (caem aleatoriamente em torno da linha central) e entre o desvio padrão em torno de 1, no caso menor que 1 perto de 0,02. O gráfico 3 apresenta a tendência central e o padrão de dependência dos resíduos.

Gráfico 3 – Resíduos versus Valores ajustados



Fonte: Produção do próprio autor

O gráfico de resíduos versus valores ajustados (tendência central) encontrado mostra que os resíduos são aleatoriamente distribuídos e têm variância constante. Os pontos caem aleatoriamente em ambos os lados de zero, sem padrões reconhecíveis nos pontos.

Por fim, a equação que descreve a relação entre a resposta e os regressores é dada pela equação (56).

$$\begin{aligned} \ln(IBOV) = & 0,00738 - 8,00\ln(IPCA) - 0,2680\ln(Cambio) - \\ & -0,000363\ln(VMDNB) \end{aligned} \quad (56)$$

5 CONCLUSÃO

O trabalho teve como objetivo aplicar a Teoria de Precificação por Arbitragem (APT) para estimar o índice de retorno de mercado após uma crise financeira utilizando-se a regressão multivariada, portanto, conseguiu-se alcançar os objetivos propostos.

Com base nos dados obtidos pelo trabalho, conclui-se que a teoria de Precificação por Arbitragem (APT) pode ser modelada através da regressão multivariada, pois as medidas de desempenho da regressão apresentaram valores satisfatórios.

Considerando-se os vários fatores e os resultados alcançados, ficou evidente que essa tarefa é um processo complicado, porque há a necessidade de se modelar cada fator e cada regressor mais de uma vez.

A coleta de dados foi um processo *forward /backward*, ou seja, escolhiam-se os fatores, estes eram modelados, e, então, os dados eram analisados para depois se concluir se esses faziam ou não parte do modelo. Deste modo foram necessárias várias iterações das etapas para obtenção do modelo de regressão multivariada.

Por fim, o modelo que melhor se adaptou ao conjunto dos dados do APT foi o que considerava os fatores escolhidos IPCA diário, Câmbio e *Volume Médio Diário Negociado no Bovespa*.

Como sugestões para trabalhos futuros, sugere-se a utilização de modelo de regressão multivariado nos dados da Bolsa de Valores Norte Americana NYSE somada ao risco país, no caso o Brasil.

Também, pode-se expandir a pesquisa utilizando ferramentas, tais como, Simulação de Monte Carlo quando for estimar o valor do índice IBOVESPA para um dia a frente.

REFERÊNCIAS BIBLIOGRÁFICAS

ADVFN. **PIB Real**. PIB Brasileiro Anual. São Paulo. Disponível em:<
<https://br.advfn.com/indicadores/pib/pib-real>>. Acesso em: 15 Junho 2018.

BANCO CENTRAL DO BRASIL. **Histórico das taxas de juros**. Histórico das taxas de juros fixadas pelo Copom e evolução da taxa Selic. São Paulo. Disponível em:<
<https://www.bcb.gov.br/Pec/Copom/Port/taxaSelic.asp>>. Acesso em: 15 Junho 2018.

BASU, D.; CHAWLA, D. An empirical test of the arbitrage pricing theory - The case of indian stock market. **Global Business Review**, Los Angeles, v. 13, n. 3, p. 421-432, 2012.

BILSON, C. et al. Selecting macroeconomic variables as explanatory factors of emerging stock market returns. **Pacific-Basin Finance Journal**, Sydney, v. 9, p. 401-426, 2001.

CHEN, N. F. Roll and Ross: Economic forces and the stock market. **Journal of Business**, Chicago, v. 59, n. 3, 1986.

DHRYME, P.; Friend, I., GULTEKIN, B. A critical reexamination of the empirical evidence on the Arbitrage Pricing Theory. **The Journal of Finance**, Chicago, v. 39, n.2, p. 323-346, 1984.

GARCIA, R.; BONOMO, M. Tests of conditional asset pricing models in the brazilian stock market. **Journal of International Money and Finance**, v. 20, p. 71-90, 2001.

IBGE, Instituto Brasileiro de Geografia e Estatística. **IPCA**. Séries históricas IPCA, Índice de preços ao consumidor. São Paulo. Disponível em:<
https://ww2.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/defaultseriesHist.shtm
 >. Acesso em: 15 Junho 2018.

LENCIONE, M. A. C. Modelos de precificação. **THESIS**, São Paulo, ano I, v. 3, p. 26-50, 2005.

MAKOWITZ, H. Portfolio selection. **The Journal of Finance**, Chicago, v. 7, n. 1, p. 77-91, 1952.

MONTGOMERY, D. C.; RUNGER, G. C. **Applied statistics and probability for engineers**. 3. ed. The United States of America: John Wiley & Sons, Inc. 2002. 822 p.

REBESCHINI, A. M.; LEAL, R. P. C. Stock fund returns and macroeconomic variables in Brazil. **Latin American Business Review**, v.17, n.2, p. 139-161, 2016.

ROLL, R.; ROSS, S. A. An empirical investigation of the Arbitrage Pricing Theory. **The Journal of Finance**, Los Angeles, v.35, n. 5, p. 1073-1103, 1980.

ROSS, S. A. The Arbitrage Theory of Capital Asset Pricing. **Journal of Economic Theory**, Los Angeles, v.13, p. 341-360, 1976.

ROSS, S. A. et al. **Princípios de administração financeira**. 2. ed. São Paulo: Editora Atlas, 2002, 530 p.

ROSS, S. A. et al. **Fundamentos de Administração Financeira**. 9. ed. São Paulo: Editora McGrawHill, 2013, 782 p.

SANTOS, M. I. C.; SILVA, M. S. Teoria de Precificação por Arbitragem: um estudo empírico no setor bancário brasileiro. **Enfoque: Reflexão Contábil**, Maringá, v. 28, n. 1, p. 70-82, 2009.

UOL. **Câmbio**. Dólar Comercial. São Paulo. Disponível em:<
<https://economia.uol.com.br/cotacoes/cambio/dolar-comercial-estados-unidos/?historico>>.
Acesso em: 10 Agosto 2018.

YAHOO! Finanças. **Índice IBOVESPA**. São Paulo. Moeda em BRL. Disponível em:<
<https://br.financas.yahoo.com/quote/%5EBVSP/history?period1=1048215600&period2=1521601200&interval=1d&filter=history&frequency=1d>>. Acesso em: 15 Junho 2018.

YLI-OLLI, P.; VIRTANEN, I. Some empirical tests of the Arbitrage Pricing Theory using transformation analysis. **Empirical Economics**, Hiedelberg, v. 17, p. 507-522, 1992.