Pós-Graduação em Ciência da Computação

Dailys Maite Aliaga Reyes

ENSAIOS DE MODELOS DE REGRESSÃO LINEAR E NÃO-LINEAR PARA DADOS SIMBÓLICOS DE TIPO INTERVALO

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2017

Dailys Maite Aliaga Reyes

ENSAIOS DE MODELOS DE REGRESSÃO LINEAR E NÃO-LINEAR PARA DADOS SIMBÓLICOS DE TIPO INTERVALO

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientadora: *Renata M. Cardoso Rodrigues de Souza*
Co-Orientador: *Francisco José de Azevêdo Cysneiros*

RECIFE
2017

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

R457e Reyes, Dailys Maite Aliaga
Ensaios de modelos de regressão linear e não-linear para dados simbólicos de tipo intervalo / Dailys Maite Aliaga Reyes. – 2017.
82 f.: il., fig., tab.

Orientadora: Renata Maria Cardoso Rodrigues de Souza.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2017.
Inclui referências e apêndices.

1. Inteligência computacional. 2. Análise de dados simbólicos. 3. Modelos de regressão. I. Souza, Renata Maria Cardoso Rodrigues de (orientadora). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2017-84

Dailys Maite Aliaga Reyes

Ensaios de modelos de regressão linear e não-linear para dados simbólicos de tipo intervalo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação

Aprovado em: 14/02/2017.

BANCA EXAMINADORA

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE
(**Orientadora**)

_____

Profa. Dra. Roberta Andrade de Araújo Fagundes
Escola Politécnica de Pernambuco/UPE

_____

Prof. Dr. Eufrásio de Andrade Lima Neto
Departamento de Estatística/UFPB

Dedico esta dissertação a todos os meus familiares, amigos e professores que me deram todo o apoio necessário para que eu chegasse até aqui. Em especial meu marido, companheiro e amigo Yaicel Gé Proenza.

Agradecimentos

Agradeço aos meus pais Alberto e Lérida, que me deram o privilégio sagrado da vida. Pelo seus esforços, carinho, dedicação e conselhos. Agradeço pelo seu apoio apesar das lágrimas que estarão sempre gravadas no meu coração quando decidi deixar meu país.

A meus avós Ramón e Rafaela, responsáveis pela formação de meu carater. Agradeço pela sua sabedoria, determinação e por me apoiar nas escolhas boas e ruins que fiz na minha vida.

Agradecimento especial a meu marido Yaicel Gé Proenza presença constante, compreensão máxima, companheiro de todas as horas e por suportar meus momentos de estresse com a galhardia de quem ama. Sem ele nada tivesse acontecido. Muito Obrigada meu amor.

A meus orientadores Renata Souza e Fransisco Cysneiros, por suas orientações e paciência na elaboração desse trabalho. Muito grata pela motivação, honestidade e por me acolher de braços abertos apesar das minhas dificuldades com o idioma português.

A minha irmã Daliannis Aliaga por me dar um sobrinho maravilhoso e por nossa amizade incondicional.

Aos meus amigos, em especial Juan González, companheiro de vida que sempre me ajuda resolvendo problemas que eu arrumo para ele, pelas conversas filosóficas sobre o futuro e por todo apoio e incentivo dedicado.

Agradeço a todos os professores do programa de pós graduação em ciências da computação, principalmente aqueles que tive contato na sala de aula, pois pude aprender muito com vocês.

A todos da UFPE e o CIn que me deram a oportunidade de estudar aqui apesar de ser estrangeira e à FACEPE pelo apoio económico.

Enfim, agradeço a todos que participaram direta ou indiretamente contribuindo para mais uma conquista alcançada até aqui. Mais uma vez, obrigada!

O único lugar aonde o sucesso vem antes do trabalho é no dicionário.

—ALBERT EINSTEIN

Resumo

A presente dissertação foi desenvolvida no marco da análise de dados simbólicos de tipo intervalo, especificamente, em modelos de regressão. Os dados simbólicos são extensões de tipos de dados clássicos. Em conjuntos de dados convencionais, os objetos são individualizados, enquanto em dados simbólicos estes são unificados por relacionamentos. Primeiramente, foi realizada uma revisão sobre dados desta natureza e das principais metodologias utilizadas para sua análise. Um novo modelo de precificação de ativos de capital (CAPM pelas siglas em inglês) foi proposto e testado para dados intervalares. A abordagem levou em conta a variação nos intervalos de preços diários em ativos de mercado, observando os preços máximos e mínimos ao invés dos preços de abertura ou fechamento que têm sido mais populares em aplicações econométricas com modelos de CAPM. Para os cálculos envolvendo intervalos de preços e retornos de ativos, as operações básicas da aritmética intervalar foram utilizadas. O modelo proposto (*i*CAPM) é uma das mais recentes aplicações CAPM intervalares, em que a estimativa do parâmetro β é um intervalo. Nesta ocasião, foi proposta uma nova interpretação para dito parâmetro em conformidade com a interpretação tradicional para o risco sistemático de ativos na área das finanças. Foram apresentados dois exemplos ilustrativos com os intervalos de preços diários da Microsoft e de Amazon, usando os retornos do mercado derivados do índice S&P500 do 01 de novembro de 2013 ao 15 de janeiro de 2015. Em conformidade com os testes estatísticos aqui realizados, os resultados da aplicação do modelo CAPM intervalar (*i*CAPM) proposto são consistentes estatísticamente, com uma explicação confiável referente aos retornos dos ativos em questão e aos retornos do mercado. Conjuntamente, foi introduzido um modelo de regressão não-linear simétrica para dados simbólicos de tipo intervalo (SNLRM-IVD), o qual ajusta um único modelo de regressão não-linear aos pontos médios (centros) e amplitudes (ranges) dos intervalos considerando a distribuição de t-Student. O desempenho do modelo foi validado através do critério estatístico da magnitude média do erro relativo, desenvolvendo experimentos no âmbito de simulações de Monte Carlo em relação a vários cenários simbólicos com $outliers$. Além do mais, o modelo proposto foi ajustado a um conjunto real de dados intervalares. A principal característica deste modelo é que proporciona estimadores não sensíveis à presença de $outliers$.

Palavras-chave: SDA. Dados Simbólicos de Tipo Intervalo. Regressão Linear e Não-linear. CAPM. Outliers.

Abstract

The present dissertation was developed within the framework of the symbolic data analysis of interval-valued type, and it is specially related to regression models. Symbolic data are extensions of classic data types. In conventional data sets, objects are individualized, while in symbolic data they are unified by relationships. At first, a deep review about the nature of this kind of data and the main methodologies used for its analysis were performed. A new capital asset pricing model (CAPM) has been proposed and tested for interval symbolic data. The approach considered the daily variation of the price ranges in market assets according to the maximum and minimum prices rather than the opening or closing prices, which have been most popular in econometric applications with CAPM models. For calculations involving price ranges and asset returns, the basic operations concerning the interval arithmetic were used. The proposed model (*i*CAPM) is one of the most recent interval CAPM applications, in which the estimate of the β-parameter is, in fact, an interval. On this occasion, a new interpretation was proposed for this parameter in accordance with the traditional interpretation for the systematic risk of the assets in the market. Two figurative examples involving the daily price ranges of Microsoft and Amazon have been presented, using the market returns from the S&P500 index in the period from November 1, 2013 to January 15, 2015. In accordance with the statistical tests performed here, the results of the application of the proposed model (*i*CAPM) are statistically consistent with a reliable explanation of the assets returns and the market returns in question. Secondly, a non-linear regression model for interval-valued data was introduced (SNLRM-IVD), which sets a single regression model to the midpoints (centers) and ranges of the intervals at once, considering the t-Student distribution. The performance of the model was validated through the statistical criterion of the average magnitude of the relative error, undergoing experiments in the scope of Monte Carlo simulations in relation to several symbolic scenarios with outliers. Finally, the proposed model was fitted to a real set of interval data. The main feature of this SNLRM-IVD is that it provides estimators that are not sensitive to the presence of outliers.

Keywords: SDA. Symbolic Interval-valued Data. Linear and Non-linear Regression. CAPM. Outliers.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Sumário

1 Introdução

1.1 Contexto e Motivação

Nas últimas décadas, muitas operações e processos têm sido digitalizados e, a cada nova transação, tais como compras com cartões de crédito, operações bancárias, ligações telefônicas, operações na bolsa de valores, entre outras, novos registros com a informação correspondente são armazenados. Estima-se que, a cada 20 meses, as empresas dobrem o volume de dados acumulados em seus computadores e dispositivos de armazenamento.

Associado a esse ritmo elevado e crescimento nas bases de dados, impera a necessidade de gerenciar e extrair informações das mesmas de forma eficiente (BOCK; DIDAY, 2012). Para fazer isso de maneira significativa, primeiro é necessário considerar o que queremos aprender com os dados e, independentemente do tamanho do conjunto de dados, pode ser de maior interesse determinar o que acontece em certas categorias ao invés dos indivíduos (BILLARD; DIDAY, 2006).

Para enfrentar os desafios do processamento de grandes conjuntos de dados, surge a Ciência de Dados ou *Data Science*. Considerada como uma ciência por si só, a Ciência de Dados é em termos gerais, a extração de conhecimento a partir dos dados, usando técnicas e teorias de muitos campos, como matemática, estatística, dados de engenharia, reconhecimento de padrões e aprendizagem, computação avançada, visualização, incerteza, modelagem, armazenamento de dados e computação de alto desempenho (DIDAY, 2016).

As técnicas estatísticas podem ser aplicadas para realizar previsões, descobrir estruturas ou associações, etc. Dentre elas cabe destacar os modelos de regressão, que ajudam a entender como determinadas variáveis influenciam outra variável, o seja, prevem o valor de uma variável dependente Y a partir das informações provenientes de um conjunto de variáveis independentes X.

Embora os métodos estatísticos tradicionais sejam muito aplicados para analisar conjuntos de dados, com o supercrescimento das tecnologias, muitas destas técnicas têm-se tornado inapropriadas para tratar dados mais complexos. Uma abordagem para atender esse problema é resumir essas bases de dados de maneira que o resultado seja um conjunto de dados tratável.

A fusão de dados não estruturados, amostras não pareadas, dados de fontes múltiplas (como mistura de dados numéricos, textuais, de imagem, redes sociais) pode ser feita em conjuntos de classes de entidades individuais, que são consideradas unidades de uma população. A descrição dessas classes pode ser intervalos, distribuições de probabilidade, sequências ponderadas, funções e similares, para expressar a variabilidade dentro da classe. Uma das vantagens desta abordagem é que os dados não estruturados e amostras não pareadas se tornam estruturados e emparelhados ao nível das classes de entidades individuais. Dessa forma, obtemos novos tipos de dados, chamados "simbólicos".

Nesse sentido, se destaca a análise de dados simbólicos (Symbolic Data Analysis (SDA) (BILLARD; DIDAY, 2006)) um paradigma que abre um vasto domínio de pesquisa e aplicações possibilitando a agregação de bases de dados clássicos em estruturas mais complexas, porém menores em tamanho. Além disso, SDA é capaz de generalizar os métodos tradicionais com dados clássicos para métodos com dados simbólicos através de desenvolvimentos exploratórios, estatísticos e representações gráficas para esses tipos de dados (BILLARD; DIDAY, 2006; DIDAY; NOIRHOMME-FRAITURE, 2008).

Os dados simbólicos podem surgir em duas situações. Em primeiro lugar, os dados originais podem ser observados e coletados no mundo real, por exemplo, acompanhando a variação de uma ação na bolsa de valores ao longo do dia, pode-se obter um intervalo de valores tendo como limites os preços mínimos e máximos da ação ao invés de todos os preços durante o dia. Em segundo lugar, os dados originais podem ser processados e transformados em listas, intervalos ou histogramas (BILLARD, 2006).

Para ilustrar essa situação poderíamos supor que um banco não estaria interessado no valor monetário na conta corrente de um indivíduo, mas na variação desse valor ao longo dos meses do ano. Neste caso, todos os dados da conta do cliente podem ser agregados produzindo dados simbólicos e, em consequência, seria extremamente atípico dizer que o valor da conta do cliente em todo o mês de maio, fosse igual a R$ 5750, quando na verdade, oscilou no intervalo de [R$ 3000; R$ 8500].

Dentre as variáveis simbólicas, as do tipo intervalo têm uma ampla gama de aplicações na análise de dados e têm sido aplicadas em áreas financeiras, em análises de trafego de rede e na mineração de dados, entre outras. Principalmente pela sua capacidade de reduzir grandes volumes de informação em um número pequeno de grupos de dados sem perda significativa da mesma, o que facilita a interpretação dos resultados. Além disso, também são relevantes no caso de aplicações com dados confidenciais em que somente permite-se conhecer o intervalo dos valores. A Tabela 1.1 mostra um exemplo de dados simbólicos de tipo intervalo com informações sobre a variação no pulso, pressão sistólica e pressão diastólica de 59 pacientes (GIL et al., 2007).

Um dos problemas que tem atraído o interesse da comunidade científica na área de SDA, é a previsão de variáveis intervalares. Embora nos últimos anos tenham sido propostas diversas extensões de técnicas e metodologias da estatística clássica para o contexto de dados simbólicos,

Tabela 1.1: Exemplo: Base de dados cardiólogica.

Observação	Pulso	Pressão Sistólica	Pressão Diastólica
1	[58; 90]	[118; 173]	[63; 102]
2	[47; 68]	[104; 161]	[71; 118]
3	[32; 114]	[131; 186]	[54; 113]
⋮	⋮	⋮	⋮
57	[40; 80]	[95; 166]	[54; 100]
58	[56; 97]	[92; 173]	[45; 107]
59	[37; 86]	[83; 140]	[45; 91]

ainda prevalecem questões não resolvidas e muito para fazer. Algumas destas questões foram a principal motivação para o desenvolvimento desta dissertação, como por exemplo, o fato de que os modelos de regressão intervalar tanto linear quanto não linear propostos pelos pesquisadores consideram dois modelos de regressão independentes para o ajuste das variáveis intervalares apesar de que poderia ser considerado um único modelo e melhorar assim a precisão do erro padrão das estimativas dos parâmetros. Além disso, os modelos de regressão não-linear para dados simbólicos de tipo intervalo tem sido pouco explorados e a regressão não-linear simétrica nesse tipo de dados até o momento não foi abordada na área de SDA

1.2 Objetivos

O objetivo principal desta dissertação é desenvolver um conjunto de soluções teóricas e aplicadas na área de regressão para dados simbólicos de tipo intervalo, cujos resultados sejam qualitativamente superiores aos métodos atualmente utilizados. Mais especificamente propomos:

1. Desenvolver um modelo de estimação de risco em dados econométricos intervalares baseado no modelo Capital Asset Pricing Model (CAPM), utilizando como variáveis os preços máximos e mínimos contidos em bancos de dados financeiros.

2. Propor um novo modelo de regressão não-linear simétrica para variáveis intervalares.

3. Realizar estudos de simulação para verificar o desempenho do modelo proposto na predição de intervalos.

4. Comparar o modelo proposto com os encontrados na literatura através do erro quadrático médio por meio de experimentos de Monte Carlo utilizando testes estatísticos t-Student.

5. Aplicar as soluções propostas neste trabalho em conjuntos de dados simbólicos de tipo intervalo em bases de dados reais para a validação experimental das mesmas.

1.3 Organização da Dissertação

Os próximos capítulos da dissertação estão organizados da seguinte forma:

Capítulo 2 - Fundamentação Teórica: Regressão intervalar simbólica. A finalidade deste capítulo é fornecer o estado da arte da abordagem simbólica em análises de dados, juntamente com a apresentação de alguns conceitos fundamentais que serão utilizados ao longo da dissertação. Também são apresentados os principais resultados de regressão linear e não-linear que existem na literatura de dados simbólicos de tipo intervalo.

Capítulo 3 - Modelo de precificação de ativos de capital intervalar. Este capítulo apresenta o estado da arte do CAPM, juntamente com os principais conceitos da aritmética intervalar. Além disso, é proposto um novo modelo CAPM para dados simbólicos de tipo intervalo usando regressão linear intervalar e são mostrados os resultados do modelo proposto quando aplicado a dados reais.

Capítulo 4 - Modelo de regressão não-linear simétrica intervalar. Este capítulo propõe um novo modelo de regressão não-linear simétrica para dados intervalares. Essa solução é inspirada na metodologia de regressão não-linear simétrica cuja principal característica é proporcionar estimadores não sensiveis à presença de *outliers*. Estudos de simulações de Monte Carlo são apresentados para avaliar a performance do modelo. Bem como aplicações a dados reais.

Capítulo 5 - Conclusões. Este capítulo apresenta os principais resultados e conclusões referentes à pesquisa realizada nesta dissertação, bem como as contribuições na área de modelos de regressão para dados simbólicos de tipo intervalo. Por fim, são apresentadas algumas perspectivas em aberto para trabalhos futuros.

2

Fundamentação Teórica: Regressão intervalar simbólica

2.1 Introdução

Este capítulo divide-se em quatro partes: inicialmente se descrevem as principais características de SDA, suas aplicações e uma revisão dos trabalhos mais relevantes desenvolvidos nesta área, os quais fundamentaram esta dissertação. A continuação, apresenta-se brevemente uma revisão da literatura de SDA relativa aos principais modelos de regressão linear para dados simbólicos de tipo intervalo, seguida de outra seção relativa aos modelos não-lineares. Por fim, na última seção se descrevem as misturas de regressão.

2.2 Análise de dados simbólicos

SDA surgiu através da influência simultânea da Análise Exploratória de Dados (BEATON; TUKEY, 1974; BOCK, 1974; GILBERT, 1990), Inteligência Artificial (BEATON; TUKEY, 1974; RUSSELL et al., 2003; LUGER, 2005) e Taxonomia Numérica (SNEATH; SOKAL, 1962). As primeiras tentativas para obter dados simbólicos a partir de dados clássicos foram realizadas por (BELSON, 1959), seguido por (MORGAN; SONQUIST, 1963) com o método Automatic Interation Detector (AID). Os primeiros algoritmos, chamados de *Conceptual Clustering*, foram apresentados por (DIDAY; SIMON, 1980; MICHALSKI; STEPP; DIDAY, 1981). Trabalhos pioneiros (DIDAY, 1987, 1989, 1991) apresentam os princípios básicos de SDA. Com isso, vários outros trabalhos foram realizados em diversas direções. A partir do final da década dos 80, SDA deixou de ser restrita a um pequeno grupo de pesquisadores para ser uma área de pesquisa bastante relevante marcada por muitas publicações e conferências (NOIRHOMME-FRAITURE; BRITO, 2011). BOCK; DIDAY (2012) apresentam de maneira sólida os conceitos de SDA e os principais métodos estatísticos desenvolvidos para manipular dados desta natureza.

Contudo, os métodos para o tratamento de dados simbólicos ainda são bastante reduzidos em número. A seguir serão comentados alguns métodos clássicos já estendidos para o tratamento

de problemas que envolvem dados simbólicos.

Na estatística descritiva por exemplo: DE CARVALHO (1995) introduziu a construção de histogramas para dados simbólicos booleanos; NOIRHOMME-FRAITURE; ROUARD (1997) apresentaram o *ZoomStar*, um método gráfico para visualizar objetos simbólicos. No caso univariado $(p = 1)$, conceitos como a média amostral, a variância amostral e a distribuição de frequência foram desenvolvidos para variáveis simbólicas (BERTRAND; GOUPIL, 2000). Posteriormente, esses conceitos foram estendidos para o caso multivariado, o seja, quando $p > 1$ (BILLARD; DIDAY, 2006; BILLARD, 2004).

Para a Análise de Componentes Principais (PCA) simbólica, CAZES et al. (1997); DOUZAL-CHOUAKRIA (1998); CHOUAKRIA; DIDAY; CAZES (1998) desenvolveram métodos para reduzir um conjunto de variáveis simbólicas de natureza intervalar $p-dimensional$ para um conjunto $s-dimensional$. O objetivo é encontrar um conjunto de s componentes que juntos expliquem ao máximo a estrutura de variação das p variáveis originais. Métodos com o mesmo propósito para dados simbólicos intervalares também foram propostos por ICHINO; YAGUCHI (1994) utilizando a métrica de Minskowsky, e por NAGABHUSHAN; GOWDA; DIDAY (1995) usando princípios de séries de Taylor. Pouco depois, LAURO; PALUMBO (2000) apresentaram técnicas novas de PCA para variáveis intervalares baseadas nos limites inferior e superior dos intervalos. Os autores asseguram que sua abordagem está em conformidade com alguns métodos desenvolvidos focados apenas no centro dos intervalos e que consideram o range como um erro de mensuração ou uma perturbação aos dados. Uma extensão mais geral de PCA para dados simbólicos intervalares foi apresentada por IRPINO (2006), incluindo a dependência temporal dos dados, por exemplo, considerando os preços de abertura e fechamento de uma ação negociada no mercado financeiro.

Outra técnica estatística muito importante desenvolvida para dados simbólicos é a análise de cluster. Neste sentido, cabe destacar os trabalhos de GOWDA; DIDAY (1991a, 1992); GURU; KIRANAGI; NAGABHUSHAN (2004), os que apresentaram as principais medidas de similaridade ou dissimilaridade para mensurar a distância entre objetos simbólicos; bem como DA SILVA (2005); BILLARD; DIDAY (2006), os que propuseram as principais medidas de distância para objetos simbólicos booleanos e modais. Para dados simbólicos de natureza intervalar, BOCK (2002) apresentou métodos de partição e visualização mediante os mapas de Kohone. CARVALHO; SOUZA (2003) desenvolveram novos métodos de cluster utilizando algoritmos do tipo nuvens dinâmicas. SOUZA; DE CARVALHO (2004) introduziram métodos de partição baseados na distância $city-block$. DE CARVALHO et al. (2006) propuseram um método dinâmico de partição baseado na distância de $Hausdorff$ e CARVALHO et al. (2007) propuseram o agrupamento de dados simbólicos intervalares também baseado na distância $Hausdorff$ adaptada. Adicionalmente, outra proposta interessante é apresentada por LIMA NETO; ANJOS (2015), os autores neste trabalho consideram um modelo de regressão para dados de tipo intervalo com estrutura de cópulas obtendo resultados relevantes quando comparado com outros modelos da literatura.

No contexto da análise fatorial para dados simbólicos a primeira abordagem foi apresentada por CAZES et al. (1997). Eles introduziram um método geométrico de classificação não supervisionado em que indivíduos são descritos por vetores de intervalos numéricos. MORINEAU et al. (1994) também apresentaram contribuições nesta área. Logo, foi proposta uma generalização da análise fatorial discriminante para dados simbólicos LAURO; VERDE; PALUMBO (2000). Uma extensão da tabela bidimensional foi proposta por GETTLER-SUMMA; PARDOUX (2000), em que os autores abordaram a análise de dados simbólicos em tabelas com três entradas, sendo o tempo ou espaço a terceira dimensão.

As árvores de decisão também foram estendidas a dados simbólicos, exemplo disso foi a generalização dos conceitos desta técnica não-paramétrica por CIAMPI et al. (2000). Adicionalmente, LLATAS; M. (2000) estudou o uso de árvores de decisão considerando que os objetos simbólicos fornecem uma amostra estratificada. Segundo eles, isto permite detectar a influência dos estratos nas regras de predição. MBALLO; DIDAY (2005) propuseram também a utilização do critério de $Kolmogorov - Smirnov$ como medida em árvores de decisão simbólica.

Os modelos de regressão também têm sido estendidos a dados simbólicos, BILLARD; DIDAY (2000) foram os primeiros a propor um modelo de regressão para dados simbólicos de natureza intervalar. A abordagem proposta por eles consiste em minimizar a soma dos quadrados dos erros para os centros dos intervalos. Dois anos mais tarde, os mesmos autores propõem uma outra abordagem ajustando dois Modelo de Regressão Linear Clássicos (MRLCs) independentes para os limites inferiores e superiores dos intervalos (BILLARD; DIDAY, 2002). BILLARD; DIDAY (2006) também incluíram variáveis explicativas, bem como estruturas hierárquicas das variáveis no âmbito da regressão simbólica. MAIA; CARVALHO (2008) estenderam o modelo de regressão L_1 para dados simbólicos intervalares, considerando a soma dos desvios absolutos como critério de minimização para a estimativa dos parâmetros. LIMA NETO; CARVALHO (2008) propuseram um novo modelo para dados intervalares baseado no centro e na amplitude dos intervalos, representação que mostrou melhor desempenho do que os métodos apresentados em BILLARD; DIDAY (2000) e (BILLARD; DIDAY, 2002). LIMA NETO; CARVALHO (2010) propuseram uma nova abordagem para ajustar o modelo de regressão linear com restrição no centro e nas amplitudes dos intervalos, a fim de assegurar a coerência matemática entre os valores previstos dos limites inferior e superior do intervalo.

No caso do modelo de regressão intervalar que assume distribuições de probabilidade para os erros, DOMINGUES; SOUZA; CYSNEIROS (2010) propuseram uma metodologia de análise de dados intervalares utilizando como base o modelo de regressão linear simétrica. Baseados na teoria do modelo linear generalizado, LIMA NETO; CORDEIRO; CARVALHO (2011) introduziram um modelo de regressão bivariada simbólica para dados intervalares. SOUZA; QUEIROZ; CYSNEIROS (2011) propuseram modelos de regressão linear logística para os limites inferiores e superiores dos intervalos, em conjunto e separadamente. Adicionalmente, FAGUNDES; DE SOUZA; CYSNEIROS (2013) introduziram um modelo de regressão robusta para a estimativa e a predição de intervalos na presença de *outliers*.

As abordagens mencionadas acima estão relacionadas a modelos de regressão linear com dados simbólicos de tipo intervalo, porém, problemas reais podem envolver aplicações não-lineares para intervalos, o que tem sido pouco explorado. Recentemente, LIMA NETO; CARVALHO (2016) apresentaram um método de regressão não-linear para dados simbólicos de tipo intervalo com distribuição normal para os erros do modelo.

Por último, no domínio de séries temporais, ARROYO; MATÉ (2006) fornecem medidas de precisão para séries temporais intervalares baseadas em distâncias de $Ichino-Yaguchi$ e $Hausdorff$. MAIA; CARVALHO; LUDERMIR (2006) apresentaram duas abordagens para a previsão de series temporais considerando variáveis simbólicas intervalares. O primeiro método ajusta dois modelos independentes (Auto-regressivo de Médias Móveis (ARMA)) sobre os centros e as amplitudes dos intervalos. O segundo método baseia-se em uma abordagem híbrida e combina um modelo ARMA com uma rede neural Perceptron Multi-Camadas (MLP) (MAIA; CARVALHO; LUDERMIR, 2008). ARROYO; MATÉ (2009), formularam novos modelos para a previsão de séries temporais simbólicas com dados de tipo histograma. Por fim, GARCÍA-ASCANIO; MATÉ (2010) promovem uma comparação entre modelos de Vetores Auto-regressivos (VAR) e MLP para dados de tipo intervalo na previsão da demanda de energia eléctrica.

2.2.1 Dados simbólicos

Os dados simbólicos são extensões de tipos de dados clássicos. Em conjuntos de dados convencionais, os objetos são individualizados, enquanto em dados simbólicos estes são unificados por relacionamentos. Em geral, dados simbólicos são mais complexos do que os dados convencionais nos seguintes aspectos (GOWDA; DIDAY, 1991b; GOWDA; RAVI, 1995):

1. Todos os objetos de um conjunto de dados simbólicos podem ou não ser definidos pelas mesmas variáveis.
2. Cada variável pode ter mais do que um valor ou mesmo um intervalo de valores.
3. Em dados simbólicos complexos, os valores que as variáveis adquirem podem incluir um ou mais objetos elementares.
4. A descrição de um dado simbólico pode depender das relações existentes entre outros dados.
5. Os valores das variáveis simbólicas podem ser tipicamente frequências de ocorrência relativa ou indicar semelhança e nível de importância de outros valores.

2.2.2 Tabela de dados simbólicos

A premissa é que o processo de obtenção de dados simbólicos deve preservar o máximo de informação possível sobre os dados e ao mesmo tempo diminuir consideravelmente o tamanho

inicial da tabela de dados. Como resultado dessa transformação são geradas novas tabelas de dados, chamadas de tabelas de dados simbólicos, onde as classes de indivíduos são descritas por pelo menos uma variável simbólica. Assim, nestas tabelas, as linhas correspondem aos indivíduos ou classes de indivíduos e as colunas são as variáveis simbólicas que descrevem esses indivíduos ou classes (DIDAY, 2016).

Considere o seguinte exemplo extraído de DIDAY; NOIRHOMME-FRAITURE (2008): em uma ilha vivem 600 pássaros, sendo 400 andorinhas, 100 avestruzes e 100 pinguins. A Tabela 2.1 consiste de 600 entradas com a informação referente à espécie, capacidade de voo e tamanho para cada um dos pássaros observados na ilha. Entretanto a Tabela 2.2 mostra, em apenas 3 entradas os dados simbólicos obtidos pelo processo de SDA agrupando as aves por espécie, na qual também foi adicionada uma nova informação referente à migração dos pássaros em diferentes períodos do ano.

Tabela 2.1: Descrição simplificada da base de dados de 600 pássaros com três variáveis.

Pássaro	Espécie	Voadora	Tamanho (cm)
1	Pinguim	{NÃO}	80
⋮	⋮	⋮	⋮
599	Andorinha	{SIM}	70
600	Avestruz	{NÃO}	125

Na Tabela 2.2, notamos que as variáveis do conjunto de dados original foram transformadas. A nova variável adicionada, que representa a migração das aves, expressa que 90% das andorinhas migram, que todos os pinguins migram e que nenhum avestruz migra.

Tabela 2.2: Descrição das três espécies de aves com conceito de migração.

Espécie	Voadora	Tamanho (cm)	Migração
Pinguim	{NÃO}	[70; 95]	[100% sim, 0% não]
Andorinha	{SIM}	[60; 85]	[90% sim, 10% não]
Avestruz	{NÃO}	[85; 160]	[0% sim, 100% não]

Outro exemplo de tabela de dados simbólicos é apresentado na Tabela 2.3, onde as linhas são indivíduos e as colunas são três variáveis simbólicas: pulso (expresso por um intervalo), marca de automóvel (expresso por um conjunto de categorias) e se faz academia (expresso por uma distribuição de pesos), também obtidas de uma agregação pelo processo SDA.

Tabela 2.3: Exemplo de uma tabela de dados simbólicos.

ID	Pulso	Marca Automóvel	Faz academia
1	[58; 90]	{Ford, Fiat}	{(3/4)sim, (1/4)não}
2	[47; 68]	{Ford, Fiat, BMW}	{(1/4)sim, (5/3)não}
3	[32; 114]	{Volkswagen, Chevrolet}	{(4/5)sim, (1/2)não}

Por outro lado, a Tabela 2.4 apresenta dados "naturalmente" simbólicos de tipo intervalo, que são os dados das temperaturas mensais mínimas e máximas registradas em 60 estações meteorológicas na China (BILLARD; DIDAY, 2006).

Tabela 2.4: Exemplo de uma tabela simbólica, com dados *naturalmente* intervalares.

Temperaturas mensais ([min; max]) - Ano 1988					
Estações	Janeiro	Fevereiro	$\cdots$	Novembro	Dezembro
AnQing	[1,8; 7,1]	[2,1; 7,2]	$\cdots$	[7,8; 17,9]	[4,3; 11,8]
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ZhoJing	[2,7; 8,4]	[2,7; 8,7]	$\cdots$	[8,2; 20]	[5,1; 13,3]

Assim, podemos resumir que cada célula de uma tabela de dados simbólicos pode conter diferentes tipos de dados, em particular (BOCK; DIDAY, 2012):

a) Um único valor quantitativo.

b) Um único valor categórico.

c) Um conjunto de valores ou categorias.

d) Um intervalo.

e) Um conjunto de valores com pesos associados.

2.2.3 Variáveis simbólicas

Como foi mencionado anteriormente, as variáveis simbólicas podem assumir, para um único indivíduo, um conjunto de categorias, intervalos, histogramas, etc. Os tipos de variáveis simbólicas mais comuns são: variáveis multi-valoradas (ordinais ou não-ordinais), variáveis de tipo intervalo e variáveis modais.

- Uma variável simbólica Y é **multi-valorada não-ordinal** se seus valores $Y(i)$ correspondem a subconjuntos finitos do domínio D : $|Y(i)| < \infty$ para todos os indivíduos $i \in E$. Por exemplo, seja E o conjunto de cidades no Brasil e Y a variável que armazena os bancos que existem nas cidades. Então pode-se ter que, $Y(Recife) = \{Bradesco, Caixa, Citibank, BB\}$

- Uma variável simbólica Y é **multi-valorada ordinal** se D suporta uma relação de ordem $\prec$, tal que, para quaisquer pares de elementos $a, b \in D$, tenhamos $a \prec b$ ou $b \prec a$. Na prática, $a \prec b$ é interpretado como a antecede b ou a é menor que b. Para quaisquer dois indivíduos $i, j \in E$, em que $a = Y(i)$ e $b = Y(j)$ são os valores observados para a variável Y, é possível definir qual deles é estritamente "melhor" do que o outro sem a utilização de qualquer escala numérica: $a \prec b$ ou $b \prec a$. Por exemplo, $Y = \{Qualidade\ do\ produto\}$ e $D = \{excelente, bom, razoável, pobre, insuficiente\}$.

- Uma variável simbólica Y é definida como **intervalar** se ela representa uma realização $\xi = [a;b] \subset R^1$, com $a \leq b$ e $\{a,b\} \in R^1$. Por exemplo, seja E um grupo de homens e $Y =$ o tempo semanal de lazer (em horas), para os indivíduos $i, j \in E$ é possível ter: $Y(i) = [3;5]$ e $Y(j) = [7;9]$.

- Todas as variáveis definidas anteriormente são também conhecidas como variáveis simbólicas *booleanas*. Existem também as variáveis modais. Uma variável simbólica Y é definida como **modal** se para cada indivíduo $i \in E$, essa variável além de apresentar um subconjunto de categorias $Y(i) \subseteq D$, apresenta também uma frequência, probabilidade ou peso $w(l)$ associado a cada categoria $l \in Y(i)$ que indica o quão frequente, típica ou relevante é a categoria l para o indivíduo i. Seja Y os cursos superiores com alunos reprovados na disciplina de estatística em x universidades. Então, para uma universidade x, $Y(x) = \textit{Ciências da computação}(0,5), \textit{Física}(0,4), \textit{Química}(0,2)$.

2.2.4 Vantagens e desvantagens da utilização da análise dados simbólicos

Os princípios que caracterizam os métodos de SDA quando comparados com as abordagens estatísticas clássicas são:

- Os objetos simbólicos são capazes de representar dados mais complexos.

- Os algoritmos de SDA permitem a geração de conceitos a partir das regras e taxonomias presentes nos dados.

- Produzem descrições gráficas que consideram a variação interna dos dados simbólicos.

As principais vantagens de utilizar dados simbólicos na descrição e análise de dados são:

- Estes apresentam um resumo do conjunto original de dados de uma maneira explicativa através de descrições baseadas em propriedades relacionadas às variáveis iniciais ou outras variáveis significativas.

- Estes podem ser facilmente transformados em uma consulta na base de dados e podem ser utilizados para propagar os conceitos extraídos entre bases de dados.

- Por serem independentes da tabela de dados inicial, os dados simbólicos são capazes de identificar qualquer indivíduo correspondente a qualquer base de dados.

- Para aplicar a análise de dados exploratória a diversas bases de dados, uma alternativa possível é a construção de objetos simbólicos a partir dos dados e a posterior aplicação dos métodos de SDA no conjunto total de objetos.

Por outro lado, a principal desvantagem envolvida na utilização de dados simbólicos é o fato da agregação dos dados, que pode acarretar a perda de informações relevantes para o domínio dos dados.

2.3 Regressão linear para dados intervalares

A análise de dados através da regressão linear está dentre as técnicas mais utilizadas para a construção de modelos para descrever o comportamento de uma variável dependente a partir de um conjunto de outras variáveis independentes. A metodologia de regressão linear é aplicada em diversas áreas de pesquisa, como a financeira, a epidemiológica, a médica, a econômica, etc. A maioria desses modelos usam a minimização da soma dos quadrados dos desvios para estimar os parâmetros. Este método dos mínimos quadrados ordinários (Ordinary Least Squares (OLS)) tem a vantagem de ser computacionalmente simples e de fornecer os melhores estimadores lineares não-viesados para os parâmetros do modelo (MONTGOMERY; PECK; VINING, 2015).

Os três modelos principais de regressão linear simbólica, sem a suposição de distribuição de probabilidades para os erros, são o método do centro, o método dos mínimos e máximos e o método do centro e da amplitude do intervalo. O processo para a estimativa dos parâmetros da regressão linear nos três métodos é baseado na minimização de critérios predeterminados usando OLS.

2.3.1 Método do centro

Este método consiste em ajustar um MRLC aos pontos médios (centros) dos intervalos e em seguida aplicar esse modelo aos limites inferior e superior dos intervalos das variáveis preditoras para prever, respectivamente, os limites inferior e superior dos intervalos da variável resposta BILLARD; DIDAY (2000). O centro é dado por:

$$\varepsilon_i^c = \frac{(\varepsilon_i^{inf} + \varepsilon_i^{sup})}{2} \tag{2.1}$$

Os limites inferior e superior da variável resposta são preditos através da aplicação do vetor de parâmetros β aos limites inferiores e superiores das variáveis regressoras. O vetor β é o mesmo para os modelos aplicados a ambos os limites inferiores e superiores.

Formalmente dito, o método do centro para variáveis simbólicas de tipo intervalo pode ser definido como: Seja $E = \{e_1, e_2, \cdots, e_n\}$ um conjunto de exemplos descritos por $p+1$ variáveis intervalares Y e $X_1, X_2, \cdots, X_n$; e seja cada exemplo $e_i \in E (i = 1, \cdots, n)$ representado por um vetor de intervalos $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, onde $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ij}, \cdots, x_{ip})$, com $x_{ij} = \xi_{ij} = [a_{ij}; b_{ij}] \in \Omega = \{[a;b] : a \leq b \text{ e } \{a,b\} \in R^1\} (j = 1, \cdots, p)$ e $y_i = \left[y_i^{inf}; y_i^{sup}\right] \in \Omega$, caracterizando os valores observados de X_j e Y. Considere Y como variável resposta e $X_1, X_2, \cdots, X_p$ como variáveis regressoras relacionadas por:

$$\begin{aligned} y_i^{inf} &= \beta_0 + \beta_1 a_{i1} + \cdots + \beta_p a_{ip} + \varepsilon_i^{inf} \\ y_i^{sup} &= \beta_0 + \beta_1 b_{i1} + \cdots + \beta_p b_{ip} + \varepsilon_i^{sup} \end{aligned} \tag{2.2}$$

Com a equação (2.1) e a equação (2.2), pode-se desenvolver o critério de minimização do método do centro:

$$\begin{aligned}\sum_{i=1}^{n}\left(\varepsilon_i^{inf}+\varepsilon_i^{sup}\right)^2 &= \sum_{i=1}^{n}(y_i^{inf}-\beta_0-\beta_1 a_{i1}-\cdots-\beta_p a_{ip} \\ &+ y_i^{sup}-\beta_0-\beta_1 b_{i1}-\cdots-\beta_p b_{ip})^2 \end{aligned} \quad (2.3)$$

que representa a soma dos quadrados dos erros dos limites inferior e superior.

Os valores dos parâmetros β que minimizam a equação (2.3) são obtidos através da diferenciação dessa equação em relação a cada elemento de β, e igualando a zero cada uma das equações obtidas resultando nas equações normais seguintes:

$$\begin{gathered} n\hat{\beta}_0+\hat{\beta}_1\textstyle\sum_{i=1}^{n}x_{i1}^c+\cdots+\hat{\beta}_p\sum_{i=1}^{n}x_{ip}^c=\sum_{i=1}^{n}y_i^c \\ \hat{\beta}_0\textstyle\sum_{i=1}^{n}x_{i1}^c+\hat{\beta}_1\sum_{i=1}^{n}(x_{i1}^c)^2+\cdots+\hat{\beta}_p\sum_{i=1}^{n}x_{ip}^c x_{i1}^c=\sum_{i=1}^{n}y_i^c x_{i1}^c \\ \vdots \\ \hat{\beta}_0\textstyle\sum_{i=1}^{n}x_{ip}^c+\hat{\beta}_1\sum_{i=1}^{n}x_{i1}^c x_{ip}^c+\cdots+\hat{\beta}_p\sum_{i=1}^{n}(x_{ip}^c)^2=\sum_{i=1}^{n}y_i^c x_{ip}^c \end{gathered}$$

em que $x_{ij}^c=(a_{ij}+b_{ij})/2$ e $y_i^c=(y_i^{inf}+y_i^{sup})/2$ são os centros dos intervalos $x_{ij}, j=1,2,\cdots,p$ e y_i, respectivamente.

Desta forma, a estimativa dos parâmetros β que minimizam a soma dos quadrados dos erros (equação (2.3)) para este método, é a solução do sistema de $p+1$ equações normais. Em notação matricial, a expressão resulta em:

$$\hat{\beta}=(\mathbf{A})^{-1}\mathbf{b}$$

em que $\mathbf{A}$ é uma matriz $(p+1)\times(p+1)$ e $\mathbf{b}$ é um vetor $(p+1)\times 1$, dados por:

$$A=\begin{pmatrix} n & \sum_i x_{i1}^c & \cdots & \sum_i x_{ip}^c \\ \sum_i x_{i1}^c & \sum_i (x_{i1}^c)^2 & \cdots & \sum_i x_{ip}^c x_{i1}^c \\ \vdots & \vdots & \vdots & \vdots \\ \sum_i x_{ip}^c & \sum_i x_{i1}^c x_{ip}^c & \cdots & \sum_i (x_{ip}^c)^2 \end{pmatrix} \text{ e } \mathbf{b}=\left(\sum_i y_i^c, \sum_i y_i^c x_i\right)$$

Ao aplicar o modelo para predizer a variável resposta Y, os valores limites para seu intervalo serão dados por:

$$\hat{y}=[\hat{y}^{inf};\hat{y}^{sup}];\ \hat{y}^{inf}=(\mathbf{x}^{inf})^T\hat{\beta} \text{ e } \hat{y}^{sup}=(\mathbf{x}^{sup})^T\hat{\beta},$$

na qual, $(\mathbf{x}^{inf})^T=(1,a_1,a_2,\cdots,a_p), (\mathbf{x}^{sup})^T=(1,b_1,b_2,\cdots,b_p)$ e $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)^T$. Tenha em conta que este método não garante que $y_i^{inf}\leq y_i^{sup}$.

2.3.2 Método dos mínimos e máximos

O método dos mínimos e máximos proposto por BILLARD; DIDAY (2002) ajusta dois MRLC independentes para os limites inferiores e superiores das variáveis simbólicas e essa é a sua grande diferença quando comparado com o método do centro. Considere o conjunto de variáveis $X_1, X_2, \cdots, X_p$ como variáveis regressoras relacionadas com uma variável resposta Y através do modelo linear:

$$\begin{aligned} y_i^{inf} &= \beta_0^{inf} + \beta_1^{inf} a_{i1} + \cdots + \beta_p^{inf} a_{ip} + \varepsilon_i^{inf} \\ y_i^{sup} &= \beta_0^{sup} + \beta_1^{sup} b_{i1} + \cdots + \beta_p^{sup} b_{ip} + \varepsilon_i^{sup} \end{aligned} \tag{2.4}$$

A partir da equação (2.4) pode-se deduzir a soma dos quadrados dos erros dos limites inferiores e superiores:

$$\begin{aligned} \sum_{i=1}^{n} \left(\varepsilon_i^{inf}\right)^2 &= \sum_{i=1}^{n} \left(y_i^{inf} - \beta_0^{inf} - \beta_1^{inf} a_{i1} - \cdots - \beta_p^{inf} a_{ip}\right)^2 \\ \sum_{i=1}^{n} \left(\varepsilon_i^{sup}\right)^2 &= \sum_{i=1}^{n} \left(y_i^{sup} - \beta_0^{sup} - \beta_1^{sup} b_{i1} - \cdots - \beta_p^{sup} b_{ip}\right)^2 \end{aligned} \tag{2.5}$$

Note que a soma dos quadrados dos resíduos de ambos limites é feita de forma independente, considerando também independentes os vetores de parâmetros β utilizados para predizer os limites da variável resposta. As equações normais do método dos mínimos e máximos podem ser obtidas diferenciando a equação (2.5) com respeito aos parâmetros β e igualando a zero o resultado.

$$\begin{gathered} n\hat{\beta}_0^{inf} + \hat{\beta}_1^{inf} \textstyle\sum_{i=1}^n a_{i1} + \cdots + \hat{\beta}_p^{inf} \sum_{i=1}^n a_{ip} = \sum_{i=1}^n y_i^{inf} \\ \hat{\beta}_0^{inf} \textstyle\sum_{i=1}^n a_{i1} + \hat{\beta}_1^{inf} \sum_{i=1}^n (a_{i1})^2 + \cdots + \hat{\beta}_p^{inf} \sum_{i=1}^n a_{ip} a_{i1} = \sum_{i=1}^n y_i^{inf} a_{i1} \\ \vdots \\ \hat{\beta}_0^{inf} \textstyle\sum_{i=1}^n a_{ip} + \hat{\beta}_1^{inf} \sum_{i=1}^n a_{i1} a_{ip} + \cdots + \hat{\beta}_p^{inf} \sum_{i=1}^n (a_{ip})^2 = \sum_{i=1}^n y_i^{inf} a_{ip} \end{gathered}$$

e,

$$\begin{gathered} n\hat{\beta}_0^{sup} + \hat{\beta}_1^{sup} \textstyle\sum_{i=1}^n b_{i1} + \cdots + \hat{\beta}_p^{sup} \sum_{i=1}^n b_{ip} = \sum_{i=1}^n y_i^{sup} \\ \hat{\beta}_0^{sup} \textstyle\sum_{i=1}^n b_{i1} + \hat{\beta}_1^{sup} \sum_{i=1}^n (b_{i1})^2 + \cdots + \hat{\beta}_p^{sup} \sum_{i=1}^n b_{ip} b_{i1} = \sum_{i=1}^n y_i^{sup} b_{i1} \\ \vdots \\ \hat{\beta}_0^{sup} \textstyle\sum_{i=1}^n b_{ip} + \hat{\beta}_1^{sup} \sum_{i=1}^n b_{i1} b_{ip} + \cdots + \hat{\beta}_p^{sup} \sum_{i=1}^n (b_{ip})^2 = \sum_{i=1}^n y_i^{sup} b_{ip} \end{gathered}$$

Os valores dos parâmetros $\beta_0^{inf}; \beta_1^{inf}, \cdots, \beta_p^{inf}$ e $\beta_0^{sup}, \beta_1^{sup}, \cdots, \beta_p^{sup}$ que minimizam a equação (2.5) podem ser escritos em notação matricial como:

$$\hat{\beta}^{inf} = \left(\hat{\beta}_0^{inf}, \hat{\beta}_1^{inf}, \cdots, \hat{\beta}_p^{inf}\right)^T = (\mathbf{A})^{-1}\mathbf{a}$$

e,

$$\hat{\beta}^{sup} = (\hat{\beta}_0^{sup}, \hat{\beta}_1^{sup}, \cdots, \hat{\beta}_p^{sup})^T = (\mathbf{B})^{-1}\mathbf{b}$$

em que $\mathbf{A}$ e $\mathbf{B}$ são matrizes de dimensão $(p+1)\times(p+1)$, $\mathbf{a}$ e $\mathbf{b}$ vetores $(p+1)\times 1$, dados por:

$$\mathbf{A} = \begin{pmatrix} n & \sum_i a_{i1} & \cdots & \sum_i a_{ip} \\ \sum_i a_{i1} & \sum_i (a_{i1})^2 & \cdots & \sum_i a_{ip}a_{i1} \\ \vdots & \vdots & \vdots & \vdots \\ \sum_i a_{ip} & \sum_i a_{i1}a_{ip} & \cdots & \sum_i (a_{ip})^2 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} n & \cdots & \sum_i b_{ip} \\ \sum_i b_{i1} & \cdots & \sum_i b_{ip}b_{i1} \\ \vdots & \vdots & \vdots \\ \sum_i b_{i1}b_{ip} & \cdots & \sum_i (b_{ip})^2 \end{pmatrix}$$

$$\mathbf{a} = \left(\sum_i y_i^{inf}, \sum_i y_i^{inf} a_{i1}, \cdots, \sum_i y_i^{inf} a_{ip}\right)^T \; e \; \mathbf{b} = \left(\sum_i y_i^{sup}, \sum_i y_i^{sup} b_{i1}, \cdots, \sum_i y_i^{sup} b_{ip}\right)^T$$

Os valores preditos para os limites inferior e superior $\hat{y} = [\hat{y}^{inf}; \hat{y}^{sup}]$ da variável resposta Y depois de aplicar o modelo são dados por:

$$\hat{y}^{inf} = (\mathbf{x}^{inf})^T \hat{\beta}^{inf} \;\text{ e }\; \hat{y}^{sup} = (\mathbf{x}^{sup})^T \hat{\beta}^{sup}$$

com

$$(\mathbf{x}^{inf})^T = (1, a_1, \cdots, a_p) \;\text{ e }\; (\mathbf{x}^{sup})^T = (1, b_1, \cdots, b_p)$$

$$\hat{\beta}^{inf} = (\hat{\beta}_0^{inf}, \hat{\beta}_1^{inf}, \cdots, \hat{\beta}_p^{inf})^T \;\text{ e }\; \hat{\beta}^{sup} = (\hat{\beta}_0^{sup}, \hat{\beta}_1^{sup}, \cdots, \hat{\beta}_p^{sup})^T$$

2.3.3 Método do centro e amplitude

Este método foi proposto por LIMA NETO; CARVALHO (2008), e estabelece as somas dos quadrados dos erros relativos aos centros e amplitudes dos intervalos como critérios de minimização independentes para a estimativa dos parâmetros, obtendo-se um modelo para o centro e outro para a amplitude. A expectativa é que com a inclusão das informações contidas nas amplitudes dos intervalos melhore a predição do modelo. O ajuste dos limites inferior e superior da variável resposta é realizado através da aplicação do vetor de parâmetros β aos centros e amplitudes das variáveis regressoras.

Sejam, y^c e x_j^c (j = 1,2,...,p) os vetores de valores relativos aos centros (c) dos intervalos das variáveis intervalares y e x_j (j = 1,2,...,p). Além disso, considere y^r e x_j^r (j = 1,2,...,p) variáveis quantitativas que assumem como valores a metade das amplitudes (r) dos intervalos das variáveis intervalares y e x_j (j = 1,2,...,p). Considere y^c e y^r como variáveis resposta e x_j^c e

x_j^r (j = 1,2,...,p) um conjunto de variáveis regressoras relacionadas por:

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_t^c \qquad i = 1,\ldots,n \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r \qquad t = 1,\ldots,n \end{aligned} \tag{2.6}$$

em que,

$$x_{ij}^c = \frac{(a_{ij}+b_{ij})}{2} \text{ e } x_{ij}^r = \frac{(b_{ij}-a_{ij})}{2}$$

$$y_i^c = \frac{(y_i^{inf}+y_i^{sup})}{2} \quad \text{e} \quad y_i^r = \frac{(y_i^{sup}-y_i^{inf})}{2}$$

Neste método, os vetores de parâmetros $(\widehat{\beta}^c)^T$ e $(\widehat{\beta}^r)^T$ são obtidos de forma independente para os centros e as amplitudes dos intervalos. Portanto, a soma dos quadrados dos erros é dada por:

$$\begin{aligned} \sum_{i=1}^{n}(\varepsilon_i^c)^2 &= \sum_{i=1}^{n}(y_i^c - \beta_0^c - \beta_1^c x_{i1}^c - \cdots - \beta_p^c x_{ip}^c)^2 \\ \sum_{i=1}^{n}(\varepsilon_i^r)^2 &= \sum_{i=1}^{n}(y_i^r - \beta_0^r - \beta_1^r x_{i1}^r - \cdots - \beta_p^r x_{ip}^r)^2 \end{aligned} \tag{2.7}$$

As equações normais do método do centro e da amplitude podem ser obtidas diferenciando a equação (2.7) com respeito aos parâmetros β e igualando a zero o resultado.

$$\begin{gathered} n\hat{\beta}_0^c + \hat{\beta}_1^c \textstyle\sum_{i=1}^n x_{i1}^c + \cdots + \hat{\beta}_p^c \sum_{i=1}^n x_{ip}^c = \sum_{i=1}^n y_i^c \\ \hat{\beta}_0^c \textstyle\sum_{i=1}^n x_{i1}^c + \hat{\beta}_1^c \sum_{i=1}^n (x_{i1}^c)^2 + \cdots + \hat{\beta}_p^c \sum_{i=1}^n x_{ip}^c x_{i1}^c = \sum_{i=1}^n y_i^c x_{i1}^c \\ \vdots \\ \hat{\beta}_0^c \textstyle\sum_{i=1}^n x_{ip}^c + \hat{\beta}_1^c \sum_{i=1}^n x_{i1}^c x_{ip}^c + \cdots + \hat{\beta}_p^c \sum_{i=1}^n (x_{ip}^c)^2 = \sum_{i=1}^n y_i^c x_{ip}^c \end{gathered}$$

e,

$$\begin{gathered} n\hat{\beta}_0^r + \hat{\beta}_1^r \textstyle\sum_{i=1}^n x_{i1}^r + \cdots + \hat{\beta}_p^r \sum_{i=1}^n x_{ip}^r = \sum_{i=1}^n y_i^r \\ \hat{\beta}_0^r \textstyle\sum_{i=1}^n x_{i1}^r + \hat{\beta}_1^r \sum_{i=1}^n (x_{i1}^r)^2 + \cdots + \hat{\beta}_p^r \sum_{i=1}^n x_{ip}^r x_{i1}^r = \sum_{i=1}^n y_i^r x_{i1}^r \\ \vdots \\ \hat{\beta}_0^r \textstyle\sum_{i=1}^n x_{ip}^r + \hat{\beta}_1^r \sum_{i=1}^n x_{i1}^r x_{ip}^r + \cdots + \hat{\beta}_p^r \sum_{i=1}^n (x_{ip}^r)^2 = \sum_{i=1}^n y_i^r x_{ip}^r \end{gathered}$$

Os estimadores de mínimos quadrados de $\beta_0^c, \beta_1^c, \cdots, \beta_p^c$ e $\beta_0^r, \beta_1^r, \cdots, \beta_p^r$ que minimizam a equação (2.7) através do sistema de equações normais acima podem ser escritas em notação matricial por:

$$\hat{\beta}^c = \left(\hat{\beta}_0^c, \hat{\beta}_1^c, \cdots, \hat{\beta}_p^c\right)^T = (\mathbf{A})^{-1}\mathbf{a}$$

$$\hat{\beta}^r = \left(\hat{\beta}_0^r, \hat{\beta}_1^r, \cdots, \hat{\beta}_p^r\right)^T = (\mathbf{B})^{-1}\mathbf{b}$$

em que **A** e **B** são matrizes $(p+1)\times(p+1)$, **a** e **b** vetores $(p+1)\times 1$, dados por:

$$\mathbf{A} = \begin{pmatrix} n & \sum_i x_{i1}^c & \cdots & \sum_i x_{ip}^c \\ \sum_i x_{i1}^c & \sum_i (x_{i1}^c)^2 & \cdots & \sum_i x_{ip}^c x_{i1}^c \\ \vdots & \vdots & \vdots & \vdots \\ \sum_i x_{ip}^c & \sum_i (x_{i1}^c) x_{ip}^c & \cdots & \sum_i (x_{ip}^c)^2 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} n & \cdots & \sum_i x_{ip}^r \\ \sum_i x_{i1}^r & \cdots & \sum_i x_{ip}^r x_{i1}^r \\ \vdots & \vdots & \vdots \\ \sum_i (x_{i1}^r) x_{ip}^r & \cdots & \sum_i (x_{ip}^r)^2 \end{pmatrix}$$

$$\mathbf{a} = \left(\sum_i y_i^c, \sum_i y_i^c x_{i1}^c, \cdots, \sum_i y_i^c x_{ip}^c \right)^T \; e \; \mathbf{b} = \left(\sum_i y_i^r x_{i1}^r, \cdots, \sum_i y_i^r x_{ip}^r \right)^T$$

O valor $y = [y^{inf}, y^{sup}]$ é predito a partir dos valores $\hat{y}^c$ e $\hat{y}^r$ como mostrado a seguir:

$$\hat{y}^{inf} = \hat{y}^c - \hat{y}^r \quad e \quad \hat{y}^{sup} = \hat{y}^c + \hat{y}^r$$

onde, $\hat{y}^c = (\mathbf{x}^c)^T \hat{\beta}^c$, $\hat{y}^r = (\mathbf{x}^r)^T \hat{\beta}^r$, $(\mathbf{x}^c)^T = (1, x_1^c, \cdots, x_p^c)$, $(\mathbf{x}^r)^T = (1, x_1^r, \cdots, x_p^r)$, $\hat{\beta}^c = (\hat{\beta}_0^c, \hat{\beta}_1^c, \cdots, \hat{\beta}_p^c)^T$ e $\hat{\beta}^r = (\hat{\beta}_0^r, \hat{\beta}_1^r, \cdots, \hat{\beta}_p^r)^T$.

2.4 Regressão não-linear para dados intervalares

Em muitos problemas do mundo real, o estudo da relação entre variáveis assume padrões não-lineares. Ao contrário dos modelos de regressão linear para dados simbólicos de tipo intervalo, os modelos que saem da linearidade têm sido pouco explorados na área de SDA. A continuação serão descritos os principais estudos encontrados na literatura correspondente.

2.4.1 Modelo de regressão não-linear para dados intervalares

O Modelo de Regressão Não-Linear para Dados de Tipo Intervalo (NLM-IVD) é uma abordagem não-linear proposta por LIMA NETO; CARVALHO (2016). Este modelo minimiza as somas de quadrados dos erros dos centros e das amplitudes dos intervalos de forma independente. Suponha um conjunto de $i = 1 \cdots n$ observações para $p+1$ variáveis intervalares Y e $X_1, X_2, \cdots, X_p$; em que cada observação i é representada pelos vetores $\mathbf{w}_i = (\mathbf{x}_i^c, y_i^c)$ e $\mathbf{r}_i = (\mathbf{x}_i^r, y_i^r)$, com $\mathbf{x}_i^c = (x_{i1}^c, x_{i2}^c, \cdots, x_{ip}^c)$ e $\mathbf{x}_i^r = (x_{i1}^r, x_{i2}^r, \cdots, x_{ip}^r)$, sendo $x_{ij}^c = (a_{ij} + b_{ij})/2, x_{ij}^r = (b_{ij} - a_{ij})/2, y_i^c = (y_i^{inf} + y_i^{sup})/2$ e $y_i^r = (y_i^{sup} - y_i^{inf})/2$. De esta forma, os valores observados para as variáveis do modelo serão X_j^c, X_j^r, Y^c e Y^r. Além disso, considere que as variáveis respostas Y^c e Y^r estão relacionadas com as variáveis X_j^c e X_j^r através de Y^c, Y^r, X_j^c e

$X_j^r, j = 1,2,\cdots,p$, relacionadas por:

$$\begin{aligned} y_i^c &= f_c(\mathbf{x}_i^c, \theta^c) + \varepsilon_i^c \\ y_i^r &= f_r(\mathbf{x}_i^r, \theta^r) + \varepsilon_i^r \end{aligned} \tag{2.8}$$

em que, ε^c e ε^r são os erros aleatórios não correlacionados com $E(\varepsilon_i^c) = 0, E(\varepsilon_i^r) = 0, Var(\varepsilon_i^c) = \sigma_c^2, Var(\varepsilon_i^r) = \sigma_r^2, Cor(\varepsilon_1^c, \varepsilon_j^c) = 0$ e $Cor(\varepsilon_1^r, \varepsilon_j^r) = 0$ para todo $i \neq j$. ε^c e ε^r são os vetores de parâmetros desconhecidos do centro e da amplitude, ambos de dimensão $(p \times 1)$ e, $f_c(\cdot) e f_r(\cdot)$ são funções não-lineares diferenciáveis.

A soma dos quadrados dos erros para o Modelo de Regressão Não-Linear para Dados de Tipo Intervalo NLM-IVD, considerando os modelos de regressão não-linear expressos pela equação (2.8) pode ser denotada como:

$$\begin{aligned} \sum_{i=1}^{n} (\varepsilon_i^c)^2 &= \sum_{i=1}^{n} [y_i^c - f_c(\mathbf{x}_i^c, \theta^c)]^2 \\ \sum_{i=1}^{n} (\varepsilon_i^r)^2 &= \sum_{i=1}^{n} [y_i^r - f_r(\mathbf{x}_i^r, \theta^r)]^2 \end{aligned} \tag{2.9}$$

Diferenciando a equação (2.9) em relação a cada elemento dos vetores θ^c e θ^r e igualando a zero, é obtido um conjunto de $2p$ equações normais relacionadas aos modelos não-lineares responsáveis pelos centros e as amplitudes dos intervalos, respectivamente:

$$\sum_{i=1}^{n} [y_i^c - f_c(\mathbf{x}_i^c, \theta^c)] \left[\frac{\partial f_c(\mathbf{x}_i^c, \theta^c)}{\partial \theta_j^c} \right]_{\theta^c = \hat{\theta}^c} = 0, j = 1,2,\cdots,p \tag{2.10}$$

$$\sum_{i=1}^{n} [y_i^r - f_r(\mathbf{x}_i^r, \theta^r)] \left[\frac{\partial f_r(\mathbf{x}_i^r, \theta^r)}{\partial \theta_j^r} \right]_{\theta^r = \hat{\theta}^r} = 0, j = 1,2,\cdots,p \tag{2.11}$$

A resolução dos sistemas de equações expressos na equação (2.10) e a equação (2.11) pode ser extremamente difícil, por isso, os autores do modelo usam métodos iterativos para obter as estimativas dos vetores de parâmetros θ^c e θ^r. Deste modo, os valores preditos $\hat{y} = [\hat{y}^{inf}; \hat{y}^{sup}]$ de uma nova observação da variável resposta Y serão obtidos a partir de $\hat{y}^c$ e de $\hat{y}^r$ da seguinte forma:

$$\hat{y}^{inf} = min(\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r) \;\; e \;\; \hat{y}^{sup} = max(\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r) \tag{2.12}$$

em que $\hat{y}^c = f_c(\mathbf{x}_i^c, \hat{\theta}^c)$ e $\hat{y}^r = f_r(\mathbf{x}_i^r$, com $\hat{\theta}^r), (\mathbf{x}^c)^T = (x_1^c, \cdots, x_p^c), (\mathbf{x}^r)^T = (x_1^r, \cdots, x_p^r), \hat{\theta}^c = (\hat{\theta}_1^c, \cdots, \hat{\theta}_p^c)^T$ e $\hat{\theta}^r = (\hat{\theta}_1^r, \cdots, \hat{\theta}_p^r)^T$. A equação (2.12) garante a coerência matemática para o NLM-IVD, sendo que $\hat{y}^{inf} \leq \hat{y}^{sup}$.

2.5 Misturas de regressão intervalar

Outros estudos interessantes encontrados na literatura de SDA sobre regressão intervalar são as misturas de regressão linear e *kernel* apresentadas por **?**). Considerando que uma amostra intervalar contém duas subpopulações, a regressão misturada ocorre quando uma relação de regressão é conhecida e a outra é desconhecida. Segundo os autores, a mistura tem maior flexibilidade do que a análise de regressão intervalar convencional e garante que $\hat{y}^{inf} \leq \hat{y}^{sup}$.

2.5.1 Mistura centro linear + amplitude *kernel*

Este método assume uma forma paramétrica linear para o centro e uma abordagem livre para a amplitude dos intervalos e a relação de mistura pode ser descrita como:

$$E(Y/\mathbf{X}) = [E(Y^c/\mathbf{X}^c) - E(Y^r/\mathbf{X}^r), E(Y^c/\mathbf{X}^c) + E(Y^r/\mathbf{X}^r)]$$

$$= [m^c(\mathbf{X}^c) - \frac{1}{2}m^r(\mathbf{X}^r), m^c(\mathbf{X}^c) + \frac{1}{2}m^r(\mathbf{X}^r)]$$

com Y a variável resposta intervalar, $\mathbf{X} = (X_1, \cdots, X_p)^T$ o vetor de variáveis preditoras intervalares, m^r uma função desconhecida e m^c uma função conhecida dada por:

$$\hat{m}^c(\mathbf{x}^c) = (x^c)^T \boldsymbol{\beta}^c$$

em que $\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \cdots, \beta_p^c)^T$ é um vetor de dimensão $p+1$.

2.5.2 Mistura centro *kernel* + amplitude linear

Este método assume uma forma não-paramétrica para o centro e uma abordagem linear para a amplitude dos intervalos e a relação de mistura pode ser descrita como:

$$E(Y/\mathbf{X}) = [E(Y^c/\mathbf{X}^c) - E(Y^r/\mathbf{X}^r), E(Y^c/\mathbf{X}^c) + E(Y^r/\mathbf{X}^r)]$$

$$= [m^c(\mathbf{X}^c) - \frac{1}{2}m^r(\mathbf{X}^r), m^c(\mathbf{X}^c) + \frac{1}{2}m^r(\mathbf{X}^r)]$$

com Y a variável resposta intervalar, $\mathbf{X} = (X_1, \cdots, X_p)^T$ o vetor de variáveis preditoras intervalares, m^c uma função desconhecida e m^r uma função conhecida dada por:

$$\hat{m}^r(\mathbf{x}^r) = (x^r)^T \boldsymbol{\beta}^r$$

em que $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \cdots, \beta_p^r)^T$ é um vetor de dimensão $p+1$.

3

Modelo de precificação de ativos de capital intervalar

3.1 Introdução

Este capítulo introduz um novo modelo para estimar o risco sistemático no CAPM com dados simbólicos do tipo intervalo, utilizando como variáveis os preços máximo e mínimo contidos nas bases de dados financeiras para explicar os retornos dos ativos. A abordagem leva em conta as informações contidas nos intervalos diários dos preços de ativos, ao invés dos preços de abertura ou fechamento, que têm sido os mais populares para estimar a equação de regressão do modelo e também usa um único modelo para ajustar os centros e as amplitudes dos intervalos. Para os cálculos envolvendo esses intervalos, as operações básicas de aritmética intervalar são utilizadas. Além disso, propomos uma interpretação para o intervalo dos betas estimados.

3.2 Motivação

Em quase todos os bancos de dados relacionados com o mercado financeiro estão disponíveis os dados diários de abertura, fechamento, máximo e mínimo dos preços de ativos. Contudo, a análise e estimativa do risco sistemático das ações baseiam-se apenas nos preços de fechamento, existindo diversas razões que sugerem que os valores diários máximos e mínimos dos preços também sejam atendidos. Neste sentido, a revisão da literatura mais competente na área revela que as observações dos preços máximos e mínimos dos ativos fornece mais informação estatística do que a avaliação dos preços de fechamento, o que de fato torna os métodos de estimação mais eficientes e robustos.

Por exemplo, as opções exóticas de tipo *lookback* fundamentam a importância da informação contida nos preços máximos e mínimos. Este tipo de instrumento é um dos mais populares no âmbito dos mercados financeiros a nível mundial e seu cálculo depende da trajetória subjacente e dos preços máximos e mínimos alcançados ao longo da mesma (GOLDMAN; SOSIN; GATTO, 1979). Em 1980 foi estendido o estimador de PARKINSON (1980) através

da inclusão da informação dos preços máximos e mínimos, com isto, o erro quadrático médio diminuiu 30% (GARMAN; KLASS, 1980). Mais recentemente, foi constatado que o Vector Error Correction Model (VECM) baseado em previsões dos preços máximo e mínimo oferece certas vantagens sobre a maior parte dos prognósticos alternativos (CHEUNG; CHEUNG; WAN, 2008). Igualmente, os preços máximos e mínimos são componentes chaves de algumas técnicas de negociação, como por exemplo, a estratégia do canal de preços, na qual uma compra ou venda é iniciada em acordo com o comportamento recente dos preços, dependendo se o fechamento for acima ou abaixo do canal superior ou inferior construído a partir dos preços máximos e mínimos diários (EDWARDS; MAGEE; BASSETTI, 2012).

Os preços máximos e mínimos utilizados em todos estes exemplos são dados *naturalmente* simbólicos de tipo intervalo. Em consequência, a estimativa do risco sistemático do ativo através do parâmetro β no modelo CAPM, poderia se tornar mais promissora utilizando os preços máximos e mínimos como um dado simbólico de tipo intervalo. Por exemplo, a primeira tentativa de utilizar dados intervalares no CAPM para estimar o coeficiente β e para medir a sensibilidade do ativo e os retornos de mercado foi proposta por PIAMSUWANNAKIT et al. (2015). No entanto, apesar do modelo proposto ter utilizado agrupações de dados semanais, ainda obteve um valor pontual para o retorno do ativo e, portanto, o tratamento não é completamente intervalar.

3.3 Modelo clássico de Precificação de Ativos de Capital

Um dos interesses na área das finanças é compreender a relação entre o retorno e o risco envolvido na alocação de recursos (investimentos). Nesse contexto, considerando também a teoria do equilíbrio de mercado em condições de risco, o Modelo de precificação de Ativos de Capital, mais conhecido mundialmente pelas siglas em inglês CAPM, é uma ótima alternativa. Distante quatro décadas do seu aparecimento, o CAPM foi desenvolvido por SHARPE (1964); LINTNER (1965); MOSSIN (1966), que se inspiraram nos trabalhos de MARKOWITZ (1952) sobre o critério da variância média. O CAPM associa o retorno esperado pelo investidor ao risco sistemático e por isso é amplamente utilizado nas finanças por analistas de mercado em geral e corretores vinculados à bolsa de valores, em particular, para mensurar o custo do capital próprio (RUPPERT, 2004). Esta quantidade é frequentemente estimada por um modelo de regressão que estabelece uma relação linear entre y_t, o retorno de uma ação no t-ésimo período, e r_{mt}, o retorno fornecido pelo mercado medido por algum índice. O CAPM é dado por:

$$y_t - r_{ft} = \alpha + \beta(r_{mt} - r_{ft}) + \varepsilon_t, \quad t = 1, \ldots, n \tag{3.1}$$

em que, r_{ft} indica a taxa livre de risco durante o t-ésimo período, α e β são os parâmetros desconhecidos do modelo e ε_t são erros aleatórios, independentes e identicamente distribuídos como $\varepsilon_t \sim N(0, \sigma^2)$. A inclusão do intercepto α permite a possibilidade de mal especificação, ou seja, a hipótese $\alpha = 0$ indica que o preço do ativo está em conformidade com o mercado.

O parâmetro β é interpretado como o risco sistemático do ativo sob estudo e é fundamental para o cálculo do custo de capital dos fundos próprios, fator básico na avaliação de qualquer projeto ou mesmo na valorização de uma empresa. Quanto menor o valor de beta (β), menor o risco da empresa e menor o retorno esperado dos investidores, e em consequência, como última instância, o custo dos seus capitais próprios (composto por ações) também será menor. Pelo contrário, se o beta é superior a 1, o risco da empresa é maior e os investidores exigirão um retorno também superior, passando o custo dos capitais próprios a ser maior também. Estes valores dos betas dependem do intervalo de tempo usado para os cálculos dos retornos e do número de retornos utilizados na análise de regressão. Deste modo, a estimativa do valor de beta a partir de retornos mensais será diferente da estimativa baseada em retornos anuais, semestrais, bimestrais, semanais ou diários (RUPPERT, 2004).

Numerosos testes empíricos têm sido realizados no que se refere à validade do CAPM. A maior dificuldade para a validação empírica do modelo provém da sua formulação em termos de antecipações, e não de realizações. Assim, um retorno esperado nem sempre é realizado. Do ponto de vista estatístico, isso introduz um erro que deveria ser igual a zero em média, mas não exatamente zero para cada ação e cada período.

Pode-se resumir o quadro histórico do CAPM conforme demonstrado na Tabela 3.1. É importante mencionar que devem ser respeitados alguns pressupostos para o desenvolvimento de testes empíricos com o CAPM, quais sejam (COPELAND et al., 1983):

1. O intercepto α não pode ser significativamente diferente de zero.
2. O parâmetro β deve ser o único fator que explique a taxa de retorno do ativo com o risco.
3. A relação em β é linear, conforme a própia formulação do modelo.
4. A taxa de retorno da carteira do mercado deve ser maior que o ativo livre de risco quando a equação do CAPM é estimada para um longo período de tempo.

Em relação as diferentes variantes do CAPM, MERTON (1973) desenvolveu o CAPM Intertemporal (ICAPM), modelo que já procurava capturar as variações ao longo do tempo considerando que a riqueza dos investidores era completamente consumida após um determinado período, porém, isso gerava inconsistências para sua avaliação. Já o Conditional-CAPM (C-CAPM) de LEWELLEN; NAGEL (2006) tem apresentado uma performance satisfatória e, portanto, tem sido bastante apreciado. Isto, em parte, porque dito modelo considera as oscilações das variâncias e covariâncias no horizonte temporal, o que gera a quebra estrutural ou não-estacionariedade das séries, e o risco sistemático não é percebido como sendo estático. A partir de estudos anteriores, ESTRADA (2002) desenvolveu o Downside-CAPM (D-CAPM), que constitui uma generalização do modelo básico, utilizando a semivariância como métrica de dispersão dos retornos, e tendo como pressuposto o estudo da perda sistêmica. Entretanto, cabe salientar que independentemente das limitações do CAPM, este “permanece como o modelo

Tabela 3.1: Resumo histórico do CAPM: evidências empíricas e limitações.

Ano	Autor	Conclusão
1952	Markowitz	Moderna teoria de carteiras. Explora todas as possibilidades da abordagem risco - retorno (MARKOWITZ, 1952)
1963	Sharpe	Desenvolveu o modelo partindo de Markowitz (SHARPE, 1963)
1965	Lintner	Coeficiente residual não seria significativamente maior do que zero e muito menor do que o excesso de prêmio de risco (LINTNER, 1965)
1970	Malkiel e Fama	Níveis de eficiência informacional: fraca, semiforte e forte (MALKIEL; FAMA, 1970)
1972	Jensen, Black e Scholes	Carteiras. Encontram um poder explicativo de 90%. Em média, os títulos de alto risco proporcionam menor retorno do que o previsto pelo CAPM (JENSEN; BLACK; SCHOLES, 1972)
1973	Fama e Mac Beth	Verificou-se que o modelo deve ser linear e que o beta é a única medida de risco para o retorno esperado do ativo (FAMA; MACBETH, 1973)
1978	Friend, Westerfield, Granito	Reafirmam a importância do risco residual na avaliação dos ativos (FRIEND; WESTERFIELD; GRANITO, 1978)
1980	Levy	Mercado israelense. No curto prazo não há relação significativa; porém, no longo prazo explica 40% das taxas médias de retorno (LEVY, 1980)
1980	Merton	As mudanças na variação dos retornos das ações podem ser detectadas a partir das variâncias das ações passadas (MERTON, 1980)
1985	Dumontier	Mercado francês. Sete fatores comuns. Somente três possuem um prêmio de risco significativo. Índice de mercado (DUMONTIER, 1985)
1991	Fama e French	Estudaram algumas ineficiências e outras variáveis estatísticamente significativas para explicar a rentabilidade média (FAMA, 1991)
1996	Jagannathan e Wang	Encontraram fortes evidências a favor do CAPM, quando utilizaram como proxy do portfolio de mercado o índice CRSP, obtendo um poder explanatório de 30% da variação $cross-section$ do retorno de 100 carteiras (JAGANNATHAN; WANG, 1996)
2004	Ribenboim	O CAPM é efetivo no mercado em equilíbrio, e suas generalizações apresentam melhor desempenho nas economias com menor eficiência de mercado e liquidez dos ativos (RIBENBOIM, 2002)
2008	Galea, Díaz-García, Vilca	Consideram o CAPM sob Distribuições Elípticas (simétricas) (GALEA; DÍAZ-GARCÍA; VILCA, 2008)
2009	Silva e Munhoz	O CAPM nas versões condicionais seria adequado quando as variáveis macroeconômicas fossem instáveis, porquanto as generalizações do modelo poderiam capturar, de forma intertemporal, as variações nos retornos das ações devido às novas informações, o que é conhecido como $time-varying\ risk$ (SILVA; MUNHOZ, 2006)
2009	Maior e Cysneiros	Foi desenvolvida a estimação do risco sistemático em modelos com erros distribuídos na classe normal assimétrica para explicar o excesso de retorno esperado de um conjunto de ações (MAIOR; CYSNEIROS, 2009)

mais utilizado no mercado de capitais para o cálculo do retorno exigido pelos acionistas de uma empresa, de maneira a compensá-los pelo risco de seu investimento" (FAMÁ; BARROS; SILVEIRA, 2001).

3.4 Aritmética Intervalar

O amadurecimento do desenvolvimento da aritmética intervalar começou com a tese de doutorado de MOORE (1962), em que o objetivo era lidar com erros numéricos e descrever uma aritmética intervalar que estendesse à aritmética real. Desde então, milhares de trabalhos investigaram esse tema. Embora possa ser uma ideia simples, a aritmética intervalar é uma técnica muito poderosa com inúmeras aplicações em matemática, ciência da computação e engenharia. Veja algumas aplicações e implementações da aritmética intervalar em (HICKEY; JU; VAN EMDEN, 2001; DOU; ZONG; LI, 2016).

Sejam $X = [\underline{X};\overline{X}]$ e $Y = [\underline{Y};\overline{Y}]$ dois intervalos, para os quais $(\underline{X},\overline{X})$ e $(\underline{Y},\overline{Y})$ são os limites inferior e superior de X e Y, respectivamente, em termos de análise de dados simbólicos, podemos caracterizar as operações matemáticas intervalares sobre essas variáveis da seguinte maneira (OLIVEIRA; DIVERIO; CLAUDIO, 1997; MOORE; KEARFOTT; CLOUD, 2009):

$$X + Y = \left[\underline{X} + \underline{Y}; \overline{X} + \overline{Y}\right]$$

$$-X = \left[-\overline{X}; -\underline{X}\right]$$

$$X - Y = X + (-Y) = \left[\underline{X} - \overline{Y}; \overline{X} - \underline{Y}\right]$$

$$X \times Y = \left[min(\underline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}, \overline{XY});\ max(\underline{XY}, \underline{X}\overline{Y}, \overline{X}\underline{Y}, \overline{XY})\right]$$

$$\frac{1}{X} = \left[\frac{1}{\overline{X}}; \frac{1}{\underline{X}}\right] \text{ se } 0 \notin X$$

$$X \div Y = X \times \frac{1}{Y} = \left[min\left(\frac{\underline{X}}{\underline{Y}}, \frac{\overline{X}}{\overline{Y}}, \frac{\underline{X}}{\overline{Y}}, \frac{\overline{X}}{\underline{Y}}\right); max\left(\frac{\underline{X}}{\underline{Y}}, \frac{\overline{X}}{\overline{Y}}, \frac{\underline{X}}{\overline{Y}}, \frac{\overline{X}}{\underline{Y}}\right)\right] \text{ se } 0 \notin Y$$

3.5 Modelo *i*CAPM

No modelo de precificação de ativos de capital para dados intervalares (*i*CAPM) introduzido nesta seção, os valores extremos a e b para representar os intervalos $[a;b]$ serão, respectivamente, os preços mínimos e máximos diários dos ativos. Para todos os cálculos envolvendo esses intervalos, são utilizadas as operações básicas da aritmética intervalar conforme foram descritas na seção 3.3.

Como no caso do CAPM clássico, nosso primeiro passo foi calcular os intervalos de retorno. O retorno é um descritor muito importante para economistas porque quantifica em percentagem a relação entre ganho ou perda e o investimento inicial. Em conformidade, o retorno

padrão de um único ativo no tempo t (vamos chamar r_t) é calculado como:

$$y_t^c = \frac{\underline{P_{y_t}} + \overline{P_{y_t}}}{2} \text{ e } x_t^c = \frac{\underline{P_{rm_t}} + \overline{P_{rm_t}}}{2} \tag{3.2}$$

$$y_t^r = \overline{P_{y_t}} - \underline{P_{y_t}} \text{ e } x_t^r = \overline{P_{rm_t}} - \underline{P_{rm_t}} \tag{3.3}$$

Neste trabalho, propomos ajustar um único modelo de regressão linear aos centros e as amplitudes dos intervalos por meio da equação:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4}$$

em que, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4)$, $\mathbf{x_1} = (\mathbf{1}_n^T, \mathbf{0}_n^T)^T$, $\mathbf{x_2} = (\mathbf{0}_n^T, \mathbf{1}_n^T)^T$, $\mathbf{x_3} = (\mathbf{x}_c^T, \mathbf{0}_n^T)^T$, $\mathbf{x_4} = (\mathbf{0}_n^T, \mathbf{x}_r^T)^T$, $\boldsymbol{\beta} = (\alpha^c, \alpha^r, \beta^c, \beta^r)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_c^T, \boldsymbol{\varepsilon}_r^T)^T$, $\boldsymbol{\varepsilon}_c = (\varepsilon_1^c, \ldots, \varepsilon_n^c)^T$, $\boldsymbol{\varepsilon}_r = (\varepsilon_1^r, \ldots, \varepsilon_n^r)^T$, $\mathbf{x_c} = (x_1^c, \ldots, x_n^c)^T$, $\mathbf{x_r} = (x_1^r, \ldots, x_n^r)^T$,$\mathbf{Y} = (\mathbf{y}_c^T, \mathbf{y}_r^T)^T$, $\mathbf{y_c} = (y_1^c, \ldots, y_n^c)^T$, $\mathbf{y_r} = (y_1^r, \ldots, y_n^r)^T$, $\mathbf{0}_n$ e $\mathbf{1}_n$ são vetores de zeros e uns, respectivamente. As estimativas dos mínimos quadrados são a solução que minimiza $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X}^T\mathbf{Y}$$

A vantagem da abordagem proposta é a melhora na precisão do erro padrão das estimativas dos parâmetros do modelo.

Uma das questões mais importantes no mercado de capitais é a consciência do nível de risco das empresas, especialmente "o risco sistêmico (risco inevitável)" que poderia afetar o retorno das ações e poderia desempenhar um papel significativo no processo de tomada de decisão. Exemplo: a inflação reduz o poder de compra da população e, em consequência, o consumo na economia. O coeficiente β indica uma medida do risco sistemático e, é o melhor preditor linear dos retornos de ativos usando os retornos do mercado como variáveis de previsão. Portanto, uma interpretação correta desse parâmetro é muito necessária (RUPPERT, 2004). Assim, para o CAPM intervalar (*i*CAPM) introduzido neste trabalho, propomos que o risco de um ativo no mercado seja estimado através de um parâmetro β intervalar dado por:

$$\widehat{\boldsymbol{\beta}}^* = [\widehat{\beta}_a^*; \widehat{\beta}_b^*],$$

em que, $\widehat{\beta}_a^* = \widehat{\beta}^c - 1/2\,\widehat{\beta}^r$ e $\widehat{\beta}_b^* = \widehat{\beta}^c + 1/2\,\widehat{\beta}^r$, sendo os valores individuais $\widehat{\beta}^c$ e $\widehat{\beta}^r$ as estimativas de beta obtidas conforme foi descrito acima. Em especial, para o modelo proposto, sugerimos que este intervalo de valores seja interpretado da seguinte maneira:

- $\widehat{\beta}_b^* < 1 \Longrightarrow$ "Risco não agressivo".
- $\widehat{\beta}_a^* > 1 \Longrightarrow$ "Risco agressivo".
- $1 \in [\widehat{\beta}_a^*; \widehat{\beta}_b^*] \Longrightarrow$ "Risco médio".

3.6 Aplicação do *i*CAPM a dados reais

Como ilustração, são considerados os dados históricos dos preços máximos e mínimos diários de duas ações (Microsoft e Amazon) entre 01 de Novembro de 2013 ao 15 de Janeiro de 2015, disponíveis na seguinte direção eletrônica: https://www.quandl.com/data/YAHOO/. A escolha do período amostral visou atender as recomendações teóricas de que períodos muito longos podem não ser muito apropriados, uma vez que as características da empresa podem mudar ao longo do tempo (BARTHOLDY; PEARE, 2000). Outras variáveis utilizadas neste estudo foram: os intervalos dos retornos da carteira de mercado e os intervalos dos retornos do ativo livre de risco.

A carteira de mercado é a carteira que abrange todos os títulos de um mercado, porém, segundo ROLL (1977), ela não é observável, sendo, portanto, uma carteira teórica. Então, utilizou-se o índice S&P500 como proxy da carteira de mercado. Esse índice foi escolhido por conter a lista das 500 empresas mais grandes que cotizam na The New York Stock Exchange (NYSE), American Express (AMEX) e National Association of Securities Dealers Automated Quotations (NASDAQ). A vantagem deste índice está na sua construção, baseada na ponderação de ações a partir do valor de cada empresa no mercado. Além do mais, o índice S&P500 não só representa a variação no preço das ações que o compõem, mas também o impacto da distribuição dos proventos, representando o retorno que, de fato, esses títulos proporcionaram.

O ativo livre de risco é, para NAKAMURA (1998), aquele que possui retornos com desvio padrão igual a zero. Por essa razão e por ser uma recomendação praticamente unânime entre os principais autores que tratam este tema na literatura de finanças (RUPPERT, 2004; MUKHERJI, 2011), as taxas do T-bill foram usadas como os retornos livres de risco.

Para a aplicação do modelo *i*CAPM foram construídos 302 intervalos, a partir dos preços máximos e mínimos dos ativos e do índice S&P500 no período considerado. Depois foi necessário calcular os intervalos dos retornos dos ativos (Microsoft e Amazon) e os intervalos dos retornos fornecidos pelo mercado (S&P500), este último representativo da relação entre a quantidade de dinheiro ganho ou perdido como resultado de um investimento e a quantidades de dinheiro investido. Para o cálculo dos retornos foi usada a subtração intervalar da seguinte maneira:

$$R_{it} = \left[\left(\frac{P_{it} - P_{it-1}}{P_{it-1}} \right) 100 \right]$$

em que:

R_{it}: intervalo de retornos da ação no tempo t;

P_{it}: intervalo de preços da ação no final do período t;

P_{it-1}: intervalo de preços da ação no final do período $t-1$.

Após o cálculo dos intervalos dos retornos, foi calculado o intervalo do prêmio pelo risco da ação no t-ésimo período $[\underline{P}_{yt}; \overline{P}_{yt}]$, dado pela diferença entre o intervalo do retorno da ação no

t-ésimo período $[\underline{y}_t;\overline{y}_t]$ e o intervalo da taxa livre de risco durante o t-ésimo período $[\underline{r}_{ft};\overline{r}_{ft}]$. Em seguida, de forma similar, foi calculado o intervalo do prêmio pelo risco do mercado no t-ésimo período $[\underline{P}_{rmt};\overline{P}_{rmt}]$, obtido pela diferença entre o intervalo do retorno fornecido pelo mercado no t-ésimo período $[\underline{r}_{mt};\overline{r}_{mt}]$ e o intervalo da taxa livre de risco durante o t-ésimo período $[\underline{r}_{ft};\overline{r}_{ft}]$.

Utilizaram-se os intervalos dos prêmios pelo risco da ação como variável dependente e os intervalos dos prêmios pelo risco do mercado como variável independente. Assim, foram calculados os valores do centro e da amplitude de cada intervalo para a regressão intervalar pela equação (3.2) e equação (3.3) e foi aplicada a regressão usando a equação (3.4). Todos os procedimentos foram realizados com um conjunto de rotinas computacionais desenvolvidas em linguagem de computador R, ver (www.R-project.org e TEAM (2014)), das quais uma parte é apresentada no Apêndice B e as restantes estão disponíveis para os usuários interessados, mediante solicitação aos autores.

3.7 Resultados e discussão

Após o ajuste do modelo de regressão *i*CAPM usando os ativos escolhidos, foram analisados os coeficientes de determinação r^2 e r^2-ajustado. Além disso, os testes de Fisher e Student foram processados com um nível de significância de 5%.

3.7.1 Dados da Microsoft

A Figura 3.1 mostra a relação das variáveis intervalares calculadas para o ativo (Microsoft) e o índice do mercado (S&P500).

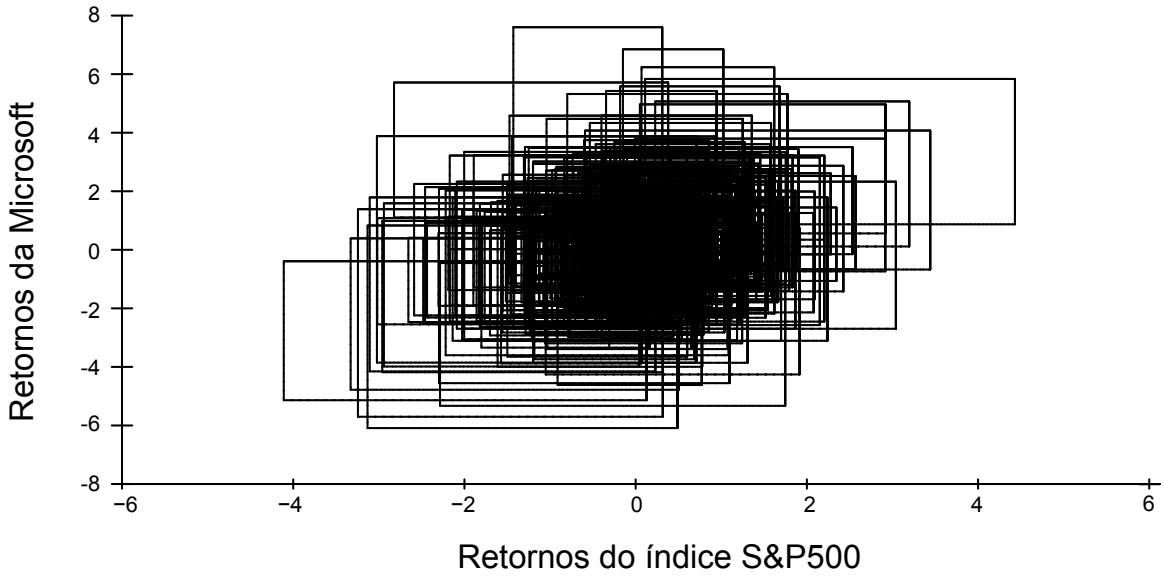

Figura 3.1: Gráfico de dispersão dos intervalos dos retornos diários da Microsoft versus intervalos dos retornos diários do índice S&P500.

A Figura 3.2 mostra a variação do prêmio de risco associado aos retornos do ativo (Microsoft) em função do prêmio de risco fornecido pelos retornos do mercado, tanto para os centros quanto para as amplitudes dos intervalos de retornos calculados. Pode-se observar que a correlação linear das variáveis relacionadas na Figura 3.2 parece mais forte para os centros dos intervalos do que para as amplitudes. Mas os olhos nem sempre são um bom juiz da intensidade de uma relação linear, por isso foi calculada uma medida numérica para suplementar o gráfico. Deste modo, o coeficiente de correlação linear ou coeficiente de Pearson computado para os centros é 0,9071 enquanto que para as amplitudes dos intervalos foi de 0,7393.

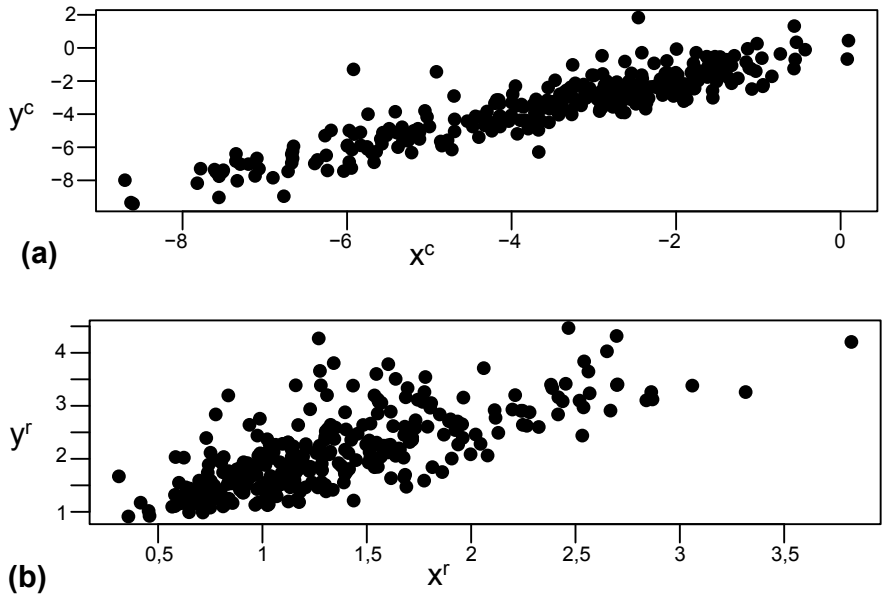

Figura 3.2: Correlação entre o prêmio de risco do ativo e o prêmio de risco do mercado, atendendo ao (a) centros e (b) amplitudes dos intervalos de retorno da Microsoft.

A Tabela 3.2 mostra as estimativas e seus respectivos erros padrão do modelo *i*CAPM ajustado para os dados da Microsoft. O coeficiente de determinação do modelo ajustado (r^2) é de $0,9531$. Este valor indica que o modelo *i*CAPM explicou 95,31% do que poderia ser explicado. O teste de adequacidade, F, do modelo indicou um p-valor de 2,2e-16, o que indica que ao nível de significância de 5% o modelo foi adequado.

Tabela 3.2: Estimativas e erros dos parâmetros do modelo de regressão ajustado *i*CAPM para os dados da Microsoft.

Parâmetro	Estimativa	Erro padrão	p-valor
α^r	0,86395	0,10281	$< 0,0001$
β^c	0,98957	0,01023	$< 0,0001$
β^r	1,09377	0,07019	$< 0,0001$

A hipótese nula $\alpha^c = 0$ foi testada, resultando não rejeitada para o centro dos intervalos o que sugere uma adequação do modelo na explicação da rendibilidade esperada. Para o ativo em análise (Microsoft), verificamos que a estimativa do parâmetro β foi positiva em ambos os

casos, o que confirma a hipótese da literatura sobre a raridade de encontrar valores negativos para este parâmetro ao analisar dados econométricos (RUPPERT, 2004).

Além disso, o intervalo estimado para o parâmetro de risco sistêmico relacionado com a Microsoft foi $\widehat{\boldsymbol{\beta}}^{*} = [0,44; 1,54]$. Por definição do *i*CAPM, o risco do ativo na carteira de mercado pode ser considerado como "risco médio". Com esses resultados, um investidor pode acreditar que na medida em que o mercado como um todo sobe, as ações do ativo em questão tendem a subir na mesma proporção, e na medida em que o mercado como um todo cai, as ações do ativo também tendem a cair na mesma proporção.

Finalmente, para verificar a intuição de que a proposta considerando um único modelo tem vantagem, foi calculado também o erro padrão das estimativas considerando dois modelos independentes para o centro e a amplitude dos intervalos. Como pode ser constatado na Tabela 3.3 o erro padrão dos parâmetros associados aos modelos independentes é maior do que o erro padrão dos parâmetros apresentado na Tabela 3.2 associados ao modelo conjunto, ainda que as estimativas sejam as mesmas.

Tabela 3.3: Estimativas e erros dos parâmetros dos modelo de regressão independentes ajustados para os centros e as amplitudes dos intervalos da Microsoft.

Parâmetro	Estimativa	Erro padrão
α^r	0,86395	0,1254
β^c	0,98957	0,1724
β^r	1,09377	0,4942

3.7.2 Dados de Amazon

A Figura 3.3 mostra a relação das variáveis intervalares calculadas para o ativo (Amazon) e o índice do mercado (S&P500).

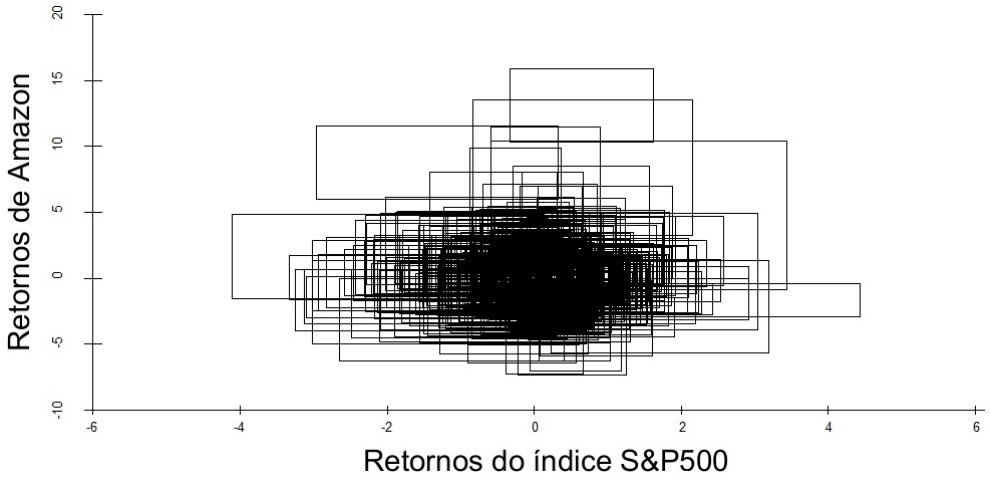

Figura 3.3: Gráfico de dispersão dos intervalos dos retornos diários de Amazon versus os intervalos dos retornos diários do índice S&P500

A Figura 3.4 mostra a variação do prêmio de risco associado aos retornos do ativo (Amazon) em função do prêmio de risco fornecido pelos retornos do mercado, tanto para os centros quanto para as amplitudes dos intervalos de retornos calculados. Pode-se observar que a correlação linear das variáveis relacionadas na Figura 3.4 é muito mais forte para os centros dos intervalos do que para as amplitudes. O coeficiente de Pearson foi calculado para medir a intensidade dessa relação linear, para os centros foi de 0.6612 enquanto que para as amplitudes dos intervalos foi de 0.2964.

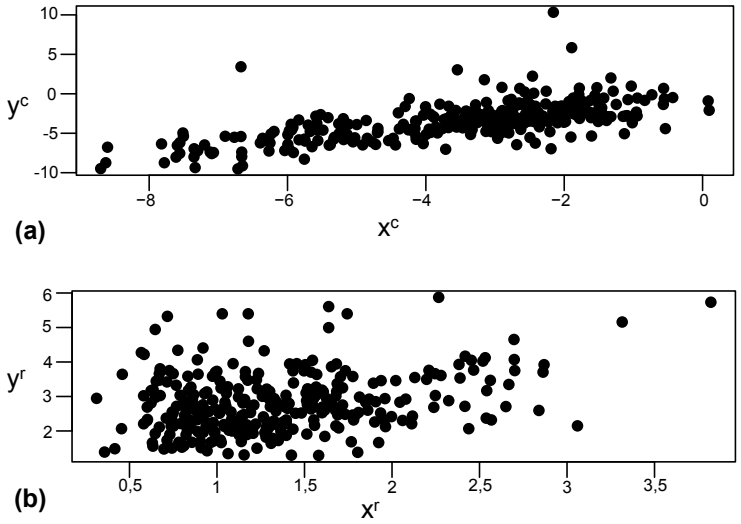

Figura 3.4: Correlação entre o prêmio de risco do ativo e o prêmio de risco do mercado, atendendo ao (a) centros e (b) amplitudes dos intervalos dos retornos de Amazon.

A Tabela 3.4 mostra as estimativas e seus respectivos erros padrão do modelo *i*CAPM ajustado para os dados de Amazon. O coeficiente de determinação do modelo ajustado (r^2) é de $0,847$, indicando que o modelo *i*CAPM explicou 84,7% do que poderia ser explicado. O teste de adequacidade, F, do modelo indicou um p-valor de 2,2e-16, o que indica que ao nível de significância de 5% o modelo foi adequado.

Tabela 3.4: Estimativas e erros dos parâmetros do modelo de regressão ajustado *i*CAPM para os dados de Amazon.

Parâmetro	Estimativa	Erro padrão	p-valor
α^r	2,20359	0,12038	$< 0,0001$
β^c	0,75685	0,02646	$< 0,0001$
β^r	0,44169	0,08218	0,00182

A hipótese nula $\alpha^c = 0$ foi testada, resultando não rejeitada para o centro dos intervalos o que sugere uma adequação do modelo na explicação da rendibilidade esperada. Verificamos que a estimativa do parâmetro β também foi positiva neste caso de estudo. Aliás, o intervalo estimado para $\widehat{\boldsymbol{\beta}}$ foi [0,54; 0,98]. Isto é, o risco do ativo na carteira de mercado pode ser considerado como "risco não agressivo", ou seja, o risco sistemático é menor do que o risco de mercado.

Como no exemplo anterior, foi calculado também o erro padrão das estimativas considerando dois modelos independentes para o centro e a amplitude dos intervalos. Como pode ser constatado na Tabela 3.5 o erro padrão dos parâmetros associados aos modelos independentes é maior do que o erro padrão dos parâmetros apresentado na Tabela 3.4 associados ao modelo conjunto, ainda que as estimativas sejam as mesmas.

Tabela 3.5: Estimativas e erros dos parâmetros dos modelo de regressão independentes ajustados para os centros e as amplitudes dos intervalos de Amazon.

Parâmetro	Estimativa	Erro padrão
α^r	2,20359	0,20655
β^c	0,75685	0,02055
β^r	0,44169	0,14101

3.8 Conclusões

Nesta seção, introduzimos um novo modelo CAPM para tratar dados simbólicos de tipo intervalo. A abordagem proposta considera os intervalos de preços máximos e mínimos diários de ativos de capital ao longo de um período ao invés dos preços de abertura ou fechamento que têm sido os mais utilizados para estimar a equação de regressão dos modelos CAPM. Também, é usado um único modelo de regressão para ajustar os centros e as amplitudes dos intervalos, melhorando a precisão do erro padrão das estimativas dos parâmetros. Para os cálculos envolvendo esses intervalos, as operações básicas da aritmética intervalar foram utilizadas. Além disso, propomos uma interpretação para a estimativa intervalar do parâmetro β e apresentamos dois exemplos ilustrativos com os intervalos de preços diários da Microsoft, de Amazon e do índice S&P500 no período de 01 de novembro de 2013 ao 15 de janeiro de 2015. Em conformidade com os testes estatísticos aqui realizados, os resultados da aplicação do modelo *i*CAPM proposto são consistentes estatísticamente, com uma explicação confiável referente aos retornos dos ativos em questão e aos retornos do mercado. Esses resultados ajudam a concluir que o modelo *i*CAPM é totalmente aplicável aos dados simbólicos de tipo intervalo.

4

Modelo de regressão não-linear simétrica intervalar

4.1 Introdução

Este capítulo introduz um novo modelo de regressão não-linear para dados simbólicos de tipo intervalo sob erros simétricos. O modelo proposto ajusta um único modelo de regressão não-linear simétrica sobre os pontos médios (centro) e amplitudes (ranges) dos intervalos. A classe de distribuições simétricas contempla distribuições de cauda pesada e de cauda leve, como exemplo, a distribuição $t-$Student e Logistica-II (DOMINGUES; SOUZA; CYSNEIROS, 2008). A principal caracteristica desta distribuição é que a estimativa dos parâmetros do modelo é menos sensível a observações *outliers* (CYSNEIROS, 2004).

4.2 Motivação

Como foi mencionado anteriormente, os modelos de regressão não-linear para dados simbólicos de tipo intervalo tem sido pouco explorado. De fato, o trabalho mais completo e recente encontrado na literatura de SDA até agora, foi apresentado por LIMA NETO; CARVALHO (2016), descrito no Capítulo 2. No entanto, a qualidade do ajuste desse modelo pode ser comprometida quando o conjunto sob investigação contém observações extremas, comumente chamadas de *outliers*.

Em conjuntos de dados clássicos, *outliers* podem ser interpretados como dados provenientes de algum erro. Porém, um pequeno número destas observações poderia não ter resultado de processos errados ou desacertos de medição. Os *outliers* podem conter informação valiosa sobre o processo que está sendo analisado e não sempre devem ser eliminados. Esta questão é ainda mais importante no caso de dados simbólicos de tipo intervalo, nos quais uma única realização intervalar representa a agregação de um extenso conjunto de medidas que, se removido poderia acarretar perda de informação importante.

Em SDA, o processo de agregação de dados é uma das principais fontes de observações

extremas em dados simbólicos de tipo intervalo. Em síntese, as descrições (conceitos) simbólicas são modeladas por processos de generalização aplicados a um conjunto de indivíduos. Pode ocorrer supergeneralização quando esses valores quantitativos são atípicos ou quando o conjunto de indivíduos se conforma de subconjuntos de diferentes distribuições, neste caso, podem surgir *outliers* (DIDAY; NOIRHOMME-FRAITURE, 2008; DOMINGUES et al., 2009).

Deste modo, são necessários modelos resistentes para estimar os parâmetros de regressão não-linear simbólica que minimizem o efeito de *outliers* em dados de natureza intervalar. Nessa direção, podemos considerar os modelos nos quais a distribuição do erro apresenta caudas mais pesadas do que a normal (tais como t-Student, logística-II, exponencial e normal contaminada, entre outras), o que pode reduzir a influência de tais observações nas estimativas dos parâmetros do modelo (PAULA; CYSNEIROS, 2009; VANEGAS; CYSNEIROS, 2010).

Várias modelos de regressão têm sido propostos baseado na distribuição simétrica para dados clássicos. Uma revisão de diferentes áreas nas quais distribuições simétricas são aplicadas pode ser encontrada em CHMIELEWSKI (1981). Além disso, uma extensão de métodos de diagnóstico para modelos não-lineares tem sido discutida em GALEA; PAULA; CYSNEIROS (2005), e uma definição geral para resíduos na classe de modelos simétricos não-lineares foi proposta por CYSNEIROS (2004). A seguir, apresenta-se o formalismo do modelo de regressão não-linear simétrica como precursor do modelo proposto neste trabalho.

4.3 Modelo clássico de regressão não-linear simétrica

Para definir um modelo de regressão com erros simétricos, suponha que a variável aleatória y (independente) tenha uma distribuição de probabilidade da classe simétrica, caracterizada por um parâmetro de locação $\mu \in \mathbb{R}$ e um parâmetro de escala $\phi > 0$. A função de densidade de probabilidade de y é definida como:

$$f(y;\mu,\phi) = \frac{1}{\sqrt{\phi}} g\left\{\frac{(y-\mu)^2}{\phi}\right\}, \qquad y \in \mathbb{R}$$

para alguma função $g(\cdot)$ denominada função geradora de densidades, com $g(u) > 0$ para $u > 0$ e $\int_0^\infty u^{-1/2} g(u)du = 1$, condição necessária para f seja uma função de distribuição de probabilidade (CHMIELEWSKI, 1981; ANDERSON; FANG, 1990; FANG; KOTZ; NG, 1990). Assim, o modelo simétrico não-linear pode ser expresso como:

$$y_i = \mu_i(\boldsymbol{\beta};\mathbf{x}_i) + \varepsilon_i, \qquad i = 1,\ldots,n \tag{4.1}$$

em que, $\mu_i = \mu_i(\boldsymbol{\beta};\mathbf{x}_i)$ é uma função não-linear contínua e diferenciável de $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$, tal que a matriz de derivadas $\mathbf{D}_\beta = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}}$ tenha posto p $(p < n)$ para todo $\boldsymbol{\beta}$, com $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^T$; $\mathbf{y} = (y_1,\ldots,y_n)^T$ é o vetor de respostas observadas; $\mathbf{x}_i = (x_{i1},\ldots,x_{ip})^T$ contém os valores de p variáveis explanatórias e $\varepsilon_i \sim S(0,\phi)$. O logaritmo da função de verossimilhança de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T,\phi)^T$

é dado por

$$L(\boldsymbol{\theta}) = -\frac{n}{2}\log\phi + \sum_{i=1}^{n}\log\{g(z_i)\}$$

em que $z_i = (y_i - \mu_i)/\sqrt{\phi}$. A função $L(\boldsymbol{\theta})$ é assumida ser regular em relação a $\boldsymbol{\beta}$ e ϕ (COX; HINKLEY, 1979). GALEA; PAULA; CYSNEIROS (2005) desenvolveram um processo iterativo para obter as estimativas de máxima verossimilhança de β e ϕ. Assim, a matriz de informações de Fisher associada para θ é da classe bloco-diagonal, ou seja, $\mathbf{K}_{\theta\theta} = \text{diag}[\mathbf{K}_{\beta\beta}, \mathbf{K}_{\phi\phi}]$, onde $\mathbf{K}_{\beta\beta} = \frac{4d_g}{\phi}(\mathbf{D}_\beta^T\mathbf{D}_\beta)$ e $\mathbf{K}_{\phi\phi} = \frac{n}{4\phi^2}(4f_g - 1)$, sendo $d_g = E\{W_g^2(u^2)u^2\}$ e $f_g = E\{W_g^2(u^2)u^4\}$, com $u \sim S(0,1)$, $W_g(u) = g'(u)/g(u)$, $g'(u) = (\partial g(u)/\partial u)$ e $v(u) = 2W_g(u)$.

4.4 Modelo de regressão não-linear simétrico para dados de tipo intervalo

A proposta apresentada neste trabalho trata-se da abordagem para dados intervalares considerando o modelo de regressão não-linear simétrico denotado por SNRM-IVD, utilizando um único modelo para o ajuste dos centros e das amplitudes dos intervalos. Pode-se constatar no Capítulo 3, que foi verificado que usar um único modelo diminui o erro padrão das estimativas quando comparado com o erro padrão das estimativas utilizando dois modelos independentes para ajustar os centros e amplitudes dos intervalos, como proposto nos modelos de regressão introduzidos até o momento na literatura. Por outro lado, as estimativas dos parâmetros do modelo Modelo de Regressão Não-Linear Simétrica para Dados de Tipo Intervalo (SNLRM-IVD) são menos sensível a presença de *outliers* do que as estimativas dos parâmetros do modelo de regressão não-linear sob erros normais, pois assume uma distribuição simétrica para os erros do modelo, por exemplo, t-Student. Além disso, essa suposição torna possível o uso de testes de hipóteses estatísticos.

Seja $\Omega = 1,\ldots,n$ uma base de dados de n objetos descritos pela variável resposta intervalar $\mathbf{Y}^I = (y_1,\ldots,y_i,\ldots,y_n)^T$ e p variáveis explanatórias intervalares $\boldsymbol{X}_1^I,\ldots,\boldsymbol{X}_j^I,\ldots,\boldsymbol{X}_p^I$ com $\boldsymbol{X}_j^I = (x_{1j},\ldots,x_{ij},\ldots,x_{nj})^T$. Cada i de Ω é representada como $(\boldsymbol{x}_i^I, y_i^I)$ com $\boldsymbol{x}_i^I = (x_{i1}^I,\ldots,x_{ip}^I)^T$, em que $x_{ij}^I = [x_{Lj}(i); x_{Uj}(i)] \in \tau = \{[a;b] : a,b \in \Re,\ a \leq b\}$ e $y_i^I = [y_L(i); y_U(i)] \in \tau$.

Neste modelo, um dado simbólico intervalar $(y^I, \boldsymbol{x}^I)$ de uma amostra ou população simbólica Ω é considerado um *outlier* quando o seu centro está a uma distância anormal dos centros dos restantes intervalos do conjunto (DOMINGUES; SOUZA; CYSNEIROS, 2010; FAGUNDES; DE SOUZA; CYSNEIROS, 2013). Assim, neste espaço, cada objeto i de Ω é representado por dois vetores $(\boldsymbol{x}_i^c, y_i^c)^T$ e $(\boldsymbol{x}_i^r, y_i^r)^T$, onde $\boldsymbol{x}_i^c = (x_{i1}^c,\ldots,x_{ip}^c)^T$ e $\boldsymbol{x}_i^r = (x_{i1}^r,\ldots,x_{ip}^r)^T$, com $x_{ij}^c = (x_{Lj}(i) + x_{Uj}(i))/2, x_{ij}^r = (x_{Uj}(i) - x_{Lj}(i))/2, y_i^c = (y_L(i) + y_U(i))/2$ e $y_i^r = (y_U(i) - y_L(i))/2$.

4.4.1 Modelo de regressão não-linear

Em relação ao vetor $\boldsymbol{Y}$ e a matriz $\boldsymbol{X}$, o modelo de regressão não-linear SNLRM-IVD pode ser escrito pela equação (4.1) como segue:

$$\boldsymbol{Y} = \mu(\boldsymbol{\beta};\boldsymbol{X}) + \boldsymbol{\varepsilon}$$

em que $\mathbf{Y} = (y_1^c,\ldots,y_n^c,y_1^r,\ldots,y_n^r)^T$ o vetor de variáveis respostas (dependentes); e considere $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3,\mathbf{X}_4)$ a matriz de variáveis preditoras (independentes), com $\mathbf{X_1} = (\mathbf{1}_n^T,\mathbf{0}_n^T)^T$, $\mathbf{X_2} = (\mathbf{0}_n^T,\mathbf{1}_n^T)^T$, $\mathbf{X_3} = (\mathbf{x}_c^T,\mathbf{0}_n^T)^T$ e $\mathbf{X_4} = (\mathbf{0}_n^T,\mathbf{x}_r^T)^T$, em que $\mathbf{0}_n$ e $\mathbf{1}_n$ são vetores de zeros e uns, respectivamente, $\mu(\boldsymbol{\beta};\boldsymbol{X})$ é uma função contínua e diferenciável de

$$\boldsymbol{\beta} = (\beta_0^c,\beta_1^c,\ldots,\beta_{p_1}^c,\beta_0^r,\beta_1^r,\ldots,\beta_{p_2}^r)^T$$

avaliada a partir dos centros e amplitudes dos intervalos conjuntamente; e $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}^c,\boldsymbol{\varepsilon}^r)^T$, com $\boldsymbol{\varepsilon}^c = (\varepsilon_1^c,\ldots,\varepsilon_n^c)^T$, $\boldsymbol{\varepsilon}^r = (\varepsilon_1^r,\ldots,\varepsilon_n^r)^T$. Adicionalmente, $\boldsymbol{\varepsilon} \sim S(0,\phi,g)$, $p = p_1 + p_2$ e $\beta_0^c,\beta_1^c,\ldots,\beta_{p_1}^c,\beta_0^r,\beta_1^r,\ldots,\beta_{p_2}^r$ são estimados pelo método da máxima verossimilhança.

4.4.2 Regra de predição

A predição dos limites inferiores e superiores dos novos intervalos realizada em função dos centros e amplitudes estimados a partir do modelo ajustado. Dado um novo objeto e seu vetor de variáveis preditoras intervalares $X^I = (x_1^I,...,x_p^I)^T$, em que cada x_j é um intervalo $x_j^I = [x_{Lj};x_{Uj}]$, sendo $x_j^c = (x_{Lj}+x_{Uj})/2 e x_j^r = (x_{Uj}-x_{Lj})/2$ os valores dos centros e amplitudes, respectivamente, do intervalo x_j. Os limites do intervalo $\hat{y} = [\hat{y}_L;\hat{y}_U]$ podem ser obtidos como segue:

$$\hat{y}_L = \hat{y}^c - \hat{y}^r/2 \text{ e } \hat{y}_U = \hat{y}^c + \hat{y}^r/2$$

em que,

$$\hat{y}^c = \hat{\beta}_0^c + \hat{\beta}_1^c x_1^c + \ldots + \hat{\beta}_p^c x_p^c$$
$$\hat{y}^r = \hat{\beta}_0^r + \hat{\beta}_1^r x_1^r + \ldots + \hat{\beta}_p^r x_p^r$$

4.5 Análise de desempenho

Experimentos de Monte Carlos foram realizados para verificar o desempenho do modelo SNLRM-IVD sob conjuntos de dados simbólicos intervalares sintéticos de complexidade hierárquica em relação à quantidade de *outliers* e também foi aplicado a um conjunto de dados reais. O objetivo foi comparar o desempenho do nosso modelo e do NLM-IVD introduzido em (LIMA NETO; CARVALHO, 2016), no que se refere à predição de dados intervalares na presença de *outliers* representativos. A precisão dos modelos foi estimada a partir da Magnitude

Média do Erro Relativo (MMRE) (**?**) através de simulações de Monte Carlo, usando o método *Hold Out* para os dados simbólicos simulados, e o método *Leave One Out* para os dados simbólicos reais. O critério da MMRE para dados simbólicos de tipo intervalo é dado por

$$MMRE = \frac{1}{2n}\sum_{i=1}^{n}\left\{\left|\frac{y_{Li}-\hat{y}_{Li}}{y_{Li}}\right| + \left|\frac{y_{Ui}-\hat{y}_{Ui}}{y_{Ui}}\right|\right\} \tag{4.2}$$

Baseado no valor médio e o desvio padrão da MMRE em cada cenário, bem como o teste estatístico t-Student para amostras pareadas a um nível de significância de 5%, o qual foi utilizado para comparar os resultados entre as abordagens, verificamos o desempenho dos modelos. Para avaliar a performance, formulamos as seguintes hipótese:
H_0: μMMRE de NLM-IVD = μMMRE de SNLRM-IVD
H_1: μMMRE de NLM-IVD $>$ μMMRE de SNLRM-IVD

Para a implementação do modelo foi utilizado o conjunto de rotinas *Elliptical*, desenvolvida por CYSNEIROS; PAULA (2005) em sua versão para R. *Elliptical* contém sub-rotinas para a análise de modelos de regressão simétrica linear e não-linear com restrições e sem restrições nos parâmetros. Alguns exemplos dos códigos implementados em R podem ser encontrados no Apêndice B, e o resto pode ser solicitado a autora.

4.5.1 Simulação de Monte Carlo

Os experimentos consistem em uma sequência de algoritmos organizados no método de simulação Monte Carlo (MC) com 1000 repetições. Inicialmente, várias relações não-lineares foram consideradas entre as variáveis respostas e explicativas para os valores dos centros e das amplitudes dos intervalos, as quais são apresentadas ao longo da discussão.

Para a primeira configuração do conjunto de dados, a equação (4.3) foi usada para gerar os centros dos intervalos, enquanto a equação (4.4) foi usada para gerar as amplitudes dos intervalos. As variáveis x_i^c e x_i^r foram obtidas de uma distribuição uniforme nos intervalos [-6; 6] e [1; 4], respectivamente.

$$y_i^c = \frac{\beta_0^c}{1+\beta_1^c\, e^{(-\beta_2^c x_i^c)}} + \varepsilon_i^c, \qquad i = 1,\ldots,n \tag{4.3}$$

$$y_i^r = \beta_0^r + e^{(-\beta_1^r x_i^r)} + \varepsilon_i^r, \qquad i = 1,\ldots,n \tag{4.4}$$

Na segunda configuração do conjunto de dados, os centros e as amplittudes das variáveis intervalares foram gerados através do modelo *Michaelis-Menten* como mostra a equação (4.5) e a equação (4.6), respectivamente. Os valores de x_i^c e x_i^r foram obtidos de uma distribuição uniforme no intervalo [5; 210]. Em ambos casos, os erros ε_i seguem uma distribuição simétrica,

$S(0,\phi)$. Nós consideramos uma distribuição t- Student com ν graus de liberdade.

$$y_i^c = \frac{\beta_1^c x_i^c}{x_i^c + \beta_0^c} + \varepsilon_i^c, \qquad i = 1,...,n \tag{4.5}$$

$$y_i^r = \frac{\beta_1^r x_i^r}{x_i^r + \beta_0^r} + \varepsilon_i^r, \qquad i = 1,...,n \tag{4.6}$$

Além disso, foram adotados três cenários de dados diferentes contendo *outliers* intervalares para avaliar o impacto dos mesmos no resultado do modelo. Estes cenários foram contaminados com diferentes percentagens de *outliers* (1%, 3% e 5%), escolhidos como os elementos inferiores (I) ou superiores (II) para a configuração 1 e aleatoriamente (III) para a configuração 2. A seguir, são apresentados os algoritmos utilizados para gerar os conjuntos de dados intervalares sintéticos para cada configuração.

4.5.1.1 Configuração 1

Foram considerados dois algoritmos (1 e 2 nos esquemas passo-a-passo) para gerar os conjuntos de dados intervalares nesta simulação MC, nos quais os cenários I e II foram simulados com 1%, 3% e 5% de *outliers* intervalares.

Algorithm 1 Simulação Monte Carlo.

Require: MC = 1000.

1: **Definir** os coeficientes na equação (4.3) e equação (4.4) são obtidos de distribuições uniformes, como segue: $\beta_0^c \sim U(1,9;2,1), \beta_1^c \sim U(2,9;3,1), \beta_2^c \sim U(0,9;1,1), \beta_0^r \sim U(0,0;0,5)$ e $\beta_1^r \sim U(0,9;1,1)$.
2: **for** j igual $1 \leq j \leq MC$ **do**
3: **Gerar** um conjunto de dados intervalares baseados no Algoritmo 2.
4: **Particionar** aleatoriamente o conjunto de intervalos gerado em conjunto de treinamento (75% dos intervalos) e conjunto de teste (25% dos intervalos).
5: **Construir** o modelo de regressão para os centros e amplitudes dos intervalos do conjunto de treinamento, seguindo os procedimentos descritos na seção 4.4.
6: **Aplicar** a regra de predição para o conjunto de teste.
7: **Calcular** as MMRE usando a equação (4.2).
8: **end for**
9: **Calcular** a média e o desvio padrão de todos os valores MMRE calculados.

A Figura 4.1 mostra o conjunto de dados simbólicos intervalares simulados para a configuração 1, na qual a Figura 4.1(a) e a Figura 4.1(b) mostram a relação entre as variáveis respostas e preditoras (y e x) para os centros e amplitudes dos intervalos de acordo como a equação (4.3) e a equação (4.4), respectivamente, enquanto a Figura 4.1(c) ilustra o conjunto de intervalos gerado inicialmente sem *outliers*.

A Figura 4.2 mostra o conjunto de dados simbólicos intervalares simulados para a configuração 1, cenário I, com $n_0 = 3\%$ (*outliers* intervalares definidos), na qual a Figura 4.2(a) e Figura 4.2(b) descrevem a relação entre as variáveis respostas e preditoras para os centros

Algorithm 2 Geração do conjunto de dados intervalar.

Require: $n = 300$.

for i igual $1 \leq i \leq n$ **do**
    Definir os erros ε_i^c e ε_i^r são obtidos de uma distribuição normal com média 0 e desvio padrão 0,05 e 0,01, respectivamente.
    Definir os valores x_i^c e x_i^r são obtidos de distribuições uniformes nos intervalos [-6; 6] e [1; 4], respectivamente.
    Calcular os valores y_i^c e y_i^r de acordo com a equação (4.3) e a equação (4.4), respectivamente.
    Definir n_0 como a quantidade de *outliers* intervalares: 1%, 3% ou 5%
    Escolher um dos cenários (Cenário I na Figura 4.2 ou Cénario II na Figura 4.3 para obter os *outliers* intervalares como segue:
    if Cenário I **then**
        Selecionar os n_0 elementos inferiores do conjunto de treinamento e substituir os valores atuais de y^c e y^r com $y^c - 3\sigma^2$ e $y^r - 3\sigma^2$, respectivamente.
    else if Cenário II **then**
        Selecionar os n_0 elementos superiores do conjunto de treinamento e substituir os valores atuais de y^c e y^r com $y^c + 3\sigma^2$ e $y^r + 3\sigma^2$, respectivamente.
    end if
end for

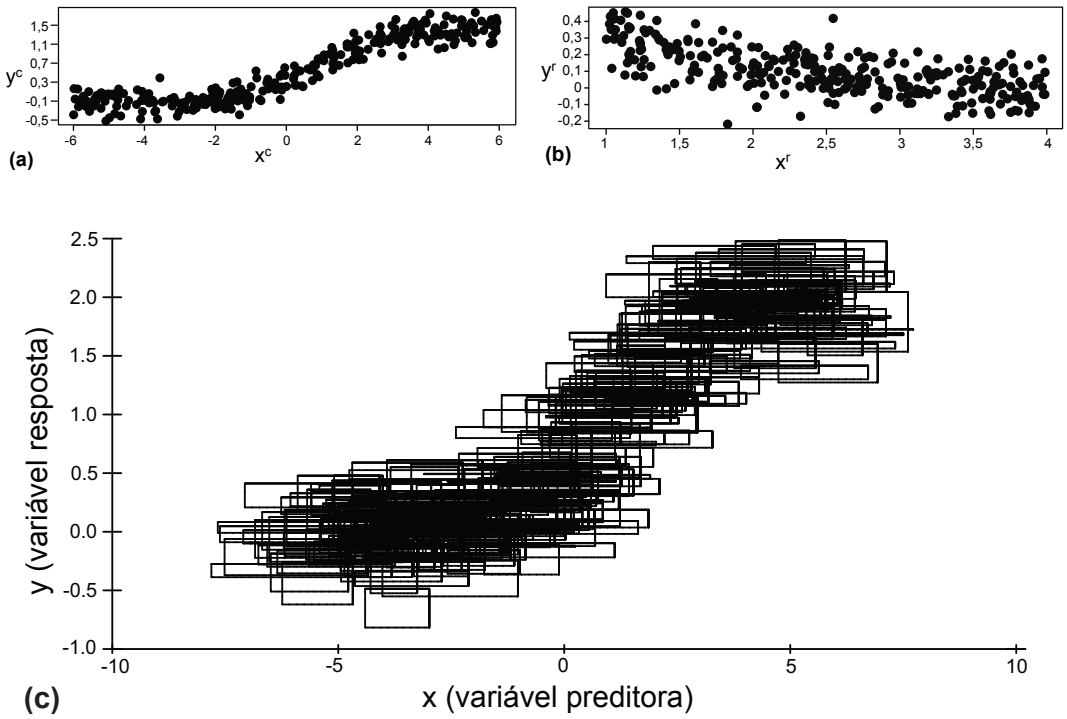

Figura 4.1: Conjunto de dados simulados para configuração 1 sem *outliers*. Os gráficos mostram a: (a) Valores dos centros dos intervalos obtidos com a equação 4.3, (b) Valores das amplitudes dos intervalos obtidos com a equação 4.4, (c) Intervalos.

e amplitudes dos intervalos, enquanto a Figura 4.2(c) representa o gráfico para os intervalos simulados.

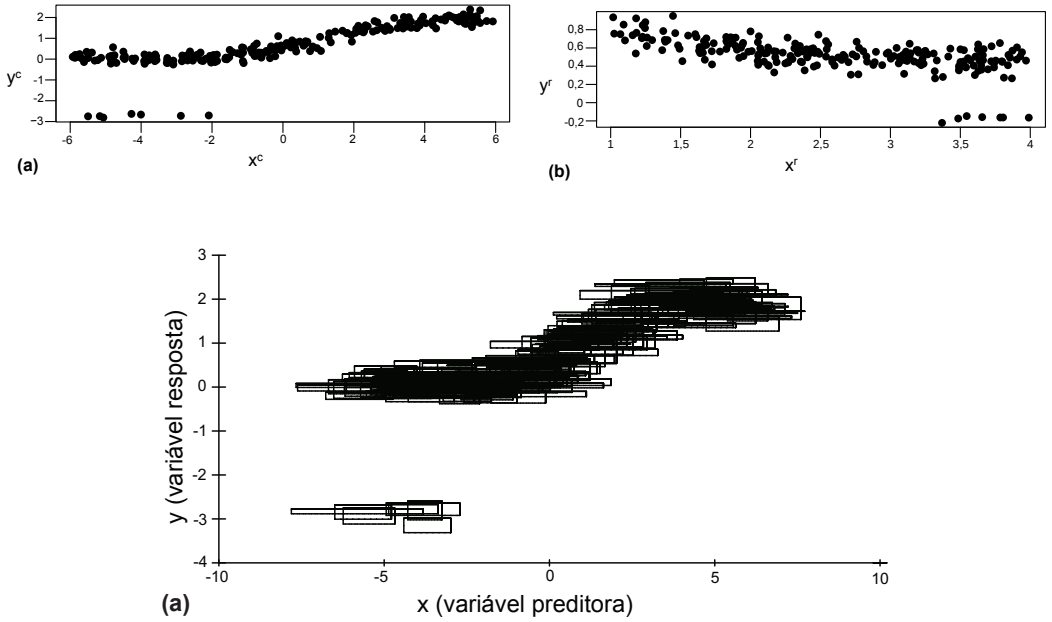

Figura 4.2: Conjunto de dados simulados para configuração 1, cenário I, com 3% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 3% de elementos inferiores como *outliers*, (b) Amplitudes dos intervalos com 3% de elementos inferiores como *outliers*, (c) Intervalos

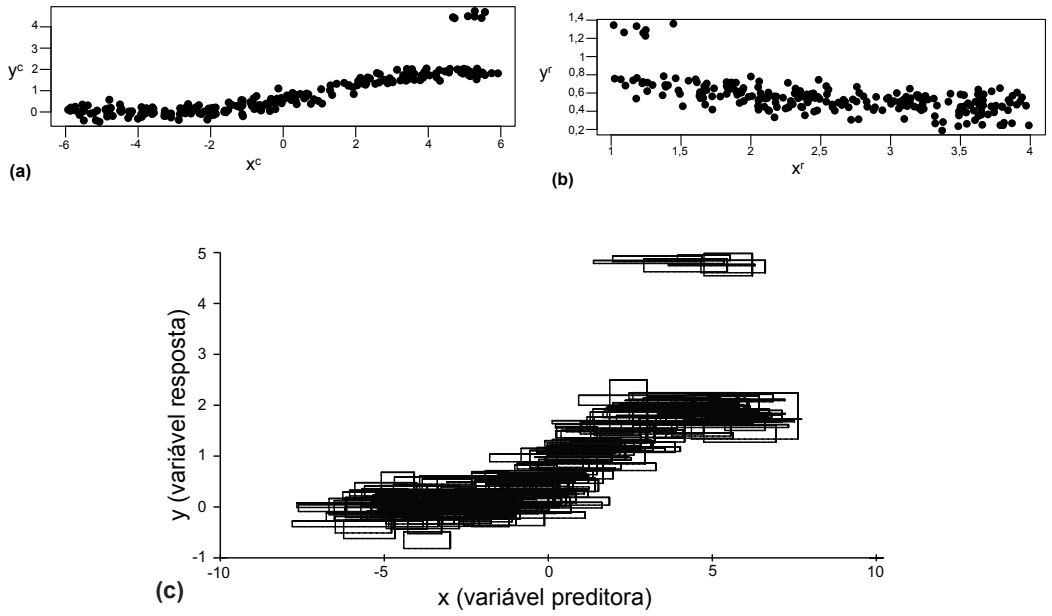

Figura 4.3: Conjunto de dados simulados para configuração 1, cenário II, com 3% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 3% de elementos superiores como *outliers*, (b) Amplitudes dos intervalos com 3% de elementos superiores como *outliers*, (c) Intervalos

A Figura 4.3 mostra o conjunto de dados simbólicos intervalares simulados para a configuração 1, cenário II, com $n_0 = 3\%$ (*outliers* intervalares definidos), na qual a Figura 4.3(a) e Figura 4.3(b) descrevem a relação entre as variáveis respostas e preditoras para os centros

e amplitudes dos intervalos, respectivamente, enquanto a Figura 4.3(c) representa o gráfico para os intervalos simulados. De maneira a facilitar a leitura do documento, os gráficos dos cenários I e II da configuração 1 com 1% e 5% de *outliers* definidos, foram colocados no Apêndice A. A Tabela 4.1 mostra os valores da média e desvio padrão das MMRE calculadas em cada experimento, além dos verdadeiros valores dos parâmetros. Pode-se apreciar que os valores das médias e desvio padrão das MMRE associados ao SNLRM-IVD são menores que os respectivos valores associados ao NLM-IVD em cada experimento feito.

Tabela 4.1: Valores das médias e desvio padrão (entre parênteses) das MMRE obtidas com os métodos NLM-IVD e SNLRM-IVD nos cenários I e II da configuração 1.

	Cenário I		Cenário II	
Outliers (%)	NLM-IVD	SNLRM-IVD	NLM-IVD	SNLRM-IVD
1%	2,0709 (1,7733)	2,0154 (1,6249)	1,0338 (0,3109)	0,9826 (0,3129)
3%	2,2765 (2,2439)	1,9860 (1,5596)	1,0883 (0,2995)	0,9835 (0,3133)
5%	2,2687 (2,2627)	1,9504 (1,862)	1,1417 (0,2764)	0,9964 (0,2959)
$\boldsymbol{\beta_0^c = 1,9;\ \beta_1^c = 3,1;\ \beta_2^c = 1,1;\ \beta_0^r = 0,4;\ \beta_1^r = 0,9}$				

Por outro lado, a Tabela 4.2 apresenta os p-valores obtidos das comparações das médias usando o teste de hipóteses t-Student pareado com um nível de significância de 5%. Esses resultados confirmam a superioridade do SNLRM-IVD sobre o NLM-IVD em ambos cenários, pois a hipóteses nula H_0 foi rejeitada com 5% de significância ($p < 0,05$). Ou seja, de acordo com esta avaliação, a abordagem SNLRM-IVD é menos sensível do que a NLM-IVD na presença de *outliers* intervalares. Além disso, o desempenho do SNLRM-IVD não é afetado pelo aumento na percentagem de *outliers*.

Tabela 4.2: p-valores do teste t-Student relacionados como os resultados dos métodos NLM-IVD e SNLRM-IVD nos cenários I e II.

Outliers (%)	Cenário I	Cenário II
1%	0,01766	0,01853
3%	0,03361	0,00784
5%	0,03985	0,00189

4.5.1.2 Configuração 2

Esta configuração foi construída sobre a mesma observações anteriores, mas usando a equação (4.5) e a equação (4.6) para obter os valores dos centros e amplitudes da variável resposta intervalar. Neste caso, apenas o cenário III foi simulado com 10% de *outliers* intervalares escolhidos aleatoriamente entre os dados. Os algoritmos 3 e 4 nos esquemas passo-a-passo, foram cosiderados pra gerar esse cenário.

A Figura 4.4 mostra o conjunto de dados simbólicos intervalares simulados para a configuração 2 com 10% de *outliers* intervalares definidos, na qual a Figura 4.4(a) e Figura 4.4(b)

Algorithm 3 Simulação Monte Carlo.

Require: $MC = 1000$.

1: **Definir** os coeficientes na equação (4.5) e equação (4.6) são obtidos de distribuições uniformes, como segue: $\beta_0^c \sim U(19;21)$, $\beta_1^c \sim U(117;123)$, $\beta_0^r \sim U(19;21)$ e $\beta_1^r \sim U(117;123)$.
2: **for** j igual $1 \leq j \leq MC$ **do**
3: **Gerar** um conjunto de dados intervalares baseados no Algoritmo 4.
4: **Particionar** aleatoriamente o conjunto de intervalos gerado em conjunto de treinamento (75% dos intervalos) e conjunto de teste (25% dos intervalos).
5: **Construir** o modelo de regressão para os centros e amplitudes dos intervalos do conjunto de treinamento, seguindo os procedimentos descritos na seção 4.4.
6: **Aplicar** a regra de predição para o conjunto de teste.
7: **Calcular** as MMRE usando a equação (4.2).
8: **end for**
9: **Calcular** a média e o desvio padrão de todos os valores MMRE calculados.

Algorithm 4 Geração do conjunto de dados intervalar.

Require: $n = 300$.

1: **for** i igual $1 \leq i \leq n$ **do**
2: **Definir** os erros ε_i^c e ε_i^r são obtidos de uma distribuição normal com média 0 e desvio padrão 5.
3: **Definir** os valores x_i^c e x_i^r são obtidos de distribuições uniformes nos intervalos [5; 210].
4: **Calcular** os valores y_i^c e y_i^r de acordo com a equação (4.5) e a equação (4.6), respectivamente.
5: **Definir** $n_0 = 10\%$ como a quantidade de *outliers* intervalares.
6: **Selecionar** os n_0 elementos aleatoriamente a partir do conjunto de treinamento.
7: **Substituir** metade (5%) dos valores de y selecionados com $y - 3\sigma^2$ e a outra metade dos valores com $y + 3\sigma^2$.
8: **end for**

descrevem a relação entre as variáveis respostas e preditoras (y e x) para os centros e amplitudes dos intervalos de acordo como a equação (4.5) e equação (4.6), respectivamente, enquanto a Figura 4.3(c) representa o gráfico para os intervalos simulados. Os resultados da comparação estatística entre os métodos NLM-IVD e SNLRM-IVD em relação ao conjunto de dados simulados no cenário III usando a equação (4.5) e equação (4.6) para ajustar o modelo, são apresentados na Tabela 4.3 os valores da média e desvio padrão das MMRE calculadas em cada experimento, além dos verdadeiros valores dos parâmetros. Neste cenário III, a média das MMRE obtidas com o SNLRM-IVD (0,0457) é menor do que a média obtida com o NLM-IVD (0,0714) e o desvio padrão também é menor para o modelo proposto (veja a Tabela 4.3). Além disso, o valor p da comparação dos resultados através do teste t-Student foi $4,065 \times 10^{-7}$, o que sugere a rejeição da hipótese nula H_0 a um nível de significância de 5%. Assim, estes resultados indicam mais uma vez que a abordagem SNLRM-IVD para dados simbólicos de tipo intervalo contaminados com dados aberrantes (*outliers*) é menos sensível a sua presença.

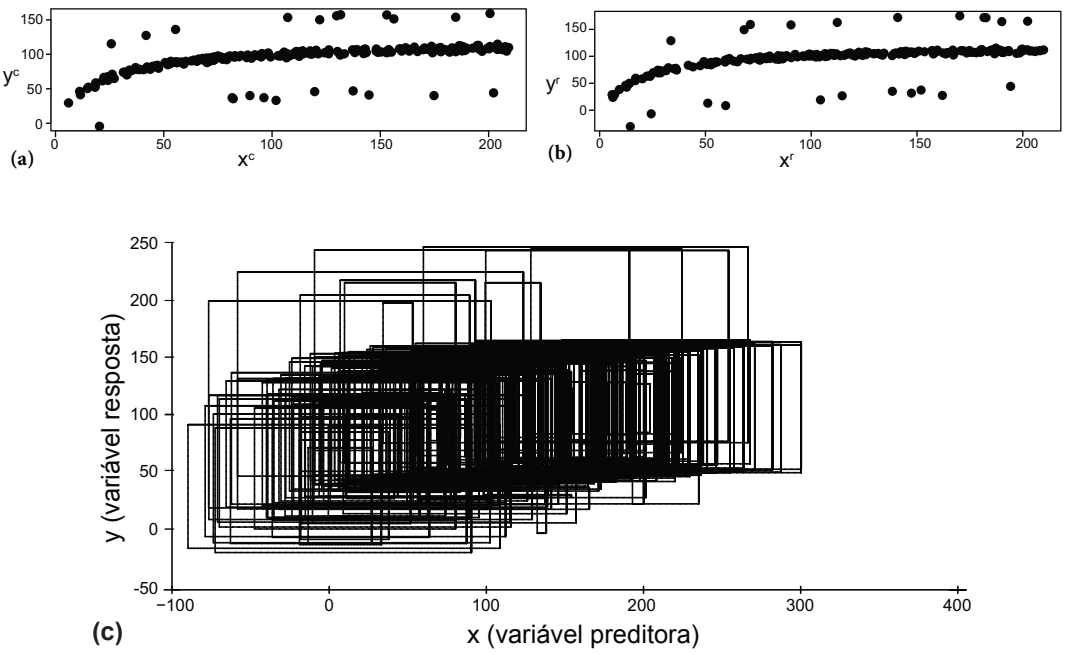

Figura 4.4: Conjunto de dados simulados para a configuração 2 com 10% de outliers intervalares. Os gráficos mostram: (a) Valores dos centros dos intervalos, (b) Valores das amplitudes dos intervalos, (c) Intervalos simulados

Tabela 4.3: Média e desvio padrão (entre parênteses) das MMRE obtidas com os métodos NLM-IVD e SNLRM-IVD para o cenário III na configuração 2.

	Cenário III	
Outliers (%)	NLM-IVD	SNLRM-IVD
10%	0,0714 (0,084)	**0,0457** (0,030)
$\boldsymbol{\beta_0^c = 19;\ \beta_1^c = 122;\ \beta_0^r = 19;\ \beta_1^r = 117}$		

4.5.2 Aplicação a dados intervalares reais

Os modelos descritos anteriormente, NLM-IVD e SNLRM-IVD, foram ajustados também a uma base real de dados simbólicos de tipo intervalo. Esses dados envolvem a informação médica disponível em BILLARD; DIDAY (2003). O exemplo, consiste em 10000 observações clássicas da consulta com o pessoal médico e informações do censo acessíveis em (http://www.census.gov), agrupadas em 42 grupos de indivíduos de acordo com a idade, diabetes e raça dos pacientes (XU, 2010). A Tabela 4.4 apresenta os intervalos do nível de glicose (Y) e da renda (X) para esses 42 grupos de indivíduos.

É importante mencionar que a escolha da função não-linear para o modelo de regressão muitas vezes depende do conhecimento prévio que se tenha sobre os dados. Neste caso, várias funções não-lineares foram testadas para ajustar o modelo de regressão com o conjunto de dados apresentado na Tabela 4.4. Para isso, consideramos o método de validação cruzada $10-fold$. Calculamos a MMRE para cada função não-linear e escolhemos a função com o menor erro de predição. Os valores iniciais foram obtidos usando um processo iterativo para a obtenção das

Tabela 4.4: Tabela intervalar com o nível de glicose (Y, em mg/dL) e renda (X, em $) de 42 grupos de indivíduos.

Grupo	Y	X	Grupo	Y	X
1	[58,5; 134]	[7268; 20369]	22	[100; 123,8]	[9180; 27951]
2	[54,6; 122]	[5337; 18897]	23	[100; 137,3]	[18195; 44771]
3	[58,8; 125]	[11400; 13421]	24	[101,1; 137,2]	[10392; 32057]
4	[59; 122]	[6582; 18200]	25	[101,1; 139,1]	[12898; 32607]
5	[67,9; 142.7]	12210; 16009]	26	[100,5; 130,7]	[8197; 20126]
6	[60,8; 121,6]	[5849; 22945]	27	[100,2; 142,4]	10906; 274542]
7	[45,9; 143,5]	[16263; 41090]	28	[101,8; 122,9]	[7215; 21993]
8	[100,9; 133,7]	[8023; 13725]	29	[101,1; 143,1]	[9913; 23472]
9	[50,1; 131,6]	[17808; 44968]	30	[101,5; 144,2]	[7327; 18700]
10	[64,8; 123,8]	[8342; 21716]	31	[100,1; 151]	[11772; 29155]
11	[58,2; 123,8]	[10928; 13000]	32	[100,3; 133,6]	[5919; 20898]
12	[70,6; 124,7]	[7751; 16711]	33	[100; 157,2]	[12237; 31505]
13	[100,5; 123,4]	[10492; 27287]	34	[100; 148,7]	[7264; 23741]
14	[77,8; 149,3]	[6298; 18958]	35	[100; 154,2]	[16368; 39687]
15	[100,4; 123,5]	[10946; 20064]	36	[101,8; 140,6]	[10643; 29799]
16	[103,6; 120,9]	[8112; 18489]	37	[100; 150,3]	[17991; 44228]
17	[100; 136,7]	[11669; 29137]	38	[100,3; 135,1]	[10136; 24325]
18	[100,2; 123,8]	[6439; 22523]	39	[100; 152,2]	[12845; 32940]
19	[100; 143,4]	[12436; 31370]	40	[102,2; 143,6]	[9591; 22516]
20	[100; 141,6]	[6836; 26498]	41	[100,1; 161,3]	[12762; 39817]
21	[100; 148,9]	[16186; 40863]	42	[102,4; 147,7]	[5257; 23259]

melhores estimativas após fornecer um chute inicial obtido por inspiração visual. A equação (4.7) e a equação (4.8) foram as melhores opções para ajustar os centros e as amplitudes dos intervalos, respectivamente.

$$y_i^c = \beta_0^c(1 - e^{(-\beta_1^c x_i^c)}) + \varepsilon_i^c, \quad i = 1, \ldots, n \tag{4.7}$$

$$y_i^r = \beta_0^r(x_i^r)^2 + \frac{\beta_1^r}{x_i^r} + \varepsilon_i^r, \quad i = 1, \ldots, n \tag{4.8}$$

A Figura 4.5 representa as linhas correspondentes às equações escolhidas para os valores dos centros e amplitudes do conjuto de dados apresentado como exemplo. Nesse gráfico, também é possivel detectar algumas observações atípicas, de tal modo, consideramos a distribuição t-Student para os erros do modelo. Um aspecto importante foi a escolha dos ν graus de liberdade (d.f) para obter uma melhor estimativa do modelo, sendo utilizado o procedimento de Akaike baseado na minimização da função AIC = $-L_\nu(\theta) + p$, estreitamente relacionada aos d.f. O modelo SNLRM-IVD com distribuição t-Student para os erros, que apresentou o menor valor AIC, foi obtido para $\nu = 5$, e portanto, estes foram os graus de liberdade escolhidos para o ajuste deste conjunto de dados reais.

A Tabela 4.5 mostra as estimativas dos parâmetros, os erros padrões e os valores de p

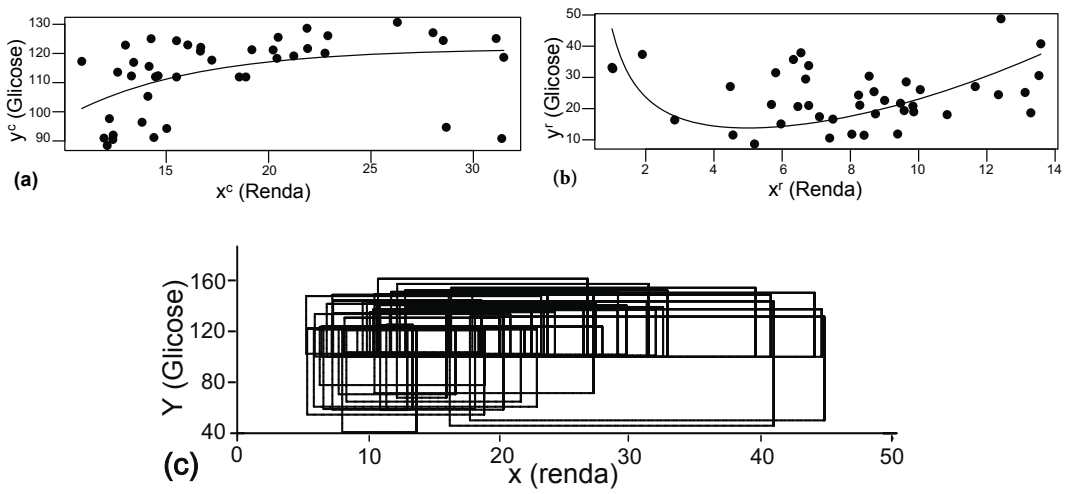

Figura 4.5: (a) Gráfico representando a linha ajustada correspondente à equação 4.7 para os valores dos centros do conjunto de dados médicos, (b) Gráfico representando a linha ajustada correspondente à equação 4.8 para os valores das amplitudes do conjunto de dados médicos, (c) Gráfico intervalar dos dados médicos

para o NLM-IVD e o SNLRM-IVD usando a distribuição t-Student com 5 d.f. Vale ressaltar que todas as estimativas dos parâmetros apresentam valores de p menores que 0,05. Neste caso, a MMRE foi obtida com o método *leave-one-out*, e o valor para SNLRM-IVD é de 4,36 é menor que o valor obtido com o NLM-IVD que é de 6,03, o que sugere que o modelo ajustado SNLRM-IVD usando a distribuição t-Student para os erros é menos sensível a presença de *outliers*. Isto corrobora que o modelo aqui proposto é uma abordagem adequada para analisar este conjunto de dados.

Tabela 4.5: Parâmetros estimados de acordo com o NLM-IVD ajustado e SNLRM-IVD usando t-Student com 5 d.f. para os dados médicos.

Modelo	Parâmetro	Estimativa	Erro padrão	p-valor
NLM-IVD	β_0^c	121,85	4,8284	$< 0{,}001$
	β_1^c	0,16	0,0303	$< 0{,}001$
	β_0^r	0,18	0,0250	$< 0{,}001$
	β_1^r	4620	8,3528	$< 0{,}001$
	ϕ	183,40	28,643	$< 0{,}001$
t-Student SNLRM-IVD	β_0^c	125,68	4,8464	$< 0{,}001$
	β_1^c	0,14	0,0231	$< 0{,}001$
	β_0^r	0,20	0,0230	$< 0{,}001$
	β_1^r	3259	7,6956	$< 0{,}001$
	ϕ	116,76	23,0656	$< 0{,}001$

4.6 Conclusões

Nesta seção, foi validado o desempenho do modelo proposto através do critério estatístico da magnitude média dos erros relativos (MMRE). Experimentos no âmbito de simulações de Monte Carlo em relação a vários cenários simbólicos com *outliers*, bem como a aplicação a um conjunto real de dados simbólicos de tipo intervalo, demonstram a robustez da abordagem simétrica com técnicas de regressão não-linear, em comparação com modelos de regressão não-linear baseados em erros normais. O SNLRM-IVD é uma metodologia versátil para investigar a relação entre variáveis simbólicas de tipo intervalo, e fornece previsões adequadas de observações intervalares mesmo na presença de *outliers*.

5

Conclusões

Nesta dissertação apresentam-se novos modelos de regressão linear e não-linear a dados simbólicos de tipo intervalo. As duas abordagens propostas foram: Modelo *i*CAPM e Modelo SNLRM-IVD.

A primeira abordagem é um novo modelo de CAPM para dados simbólicos de tipo intervalo usando regressão linear intervalar. Esse modelo associa o retorno esperado pelo investidor ao risco sistemático para mensurar o custo do capital próprio. A abordagem proposta considera os intervalos de preços máximos e mínimos diários de ativos de capital ao longo de um período ao invés dos preços de abertura ou fechamento que têm sido os mais utilizados para estimar a equação de regressão dos modelos CAPM. Também, propomos um novo conceito para classificação do ativo e apresentamos dois exemplos ilustrativos com os intervalos de preços diários da Microsoft, de Amazon e do índice S&P500 no período de 01 de novembro de 2013 ao 15 de janeiro de 2015. De acordo com os testes estatísticos aqui realizados, a aplicação do *i*CAPM com intervalos de preços, especificamente conformados a partir dos preços máximos e mínimos, é uma alternativa robusta e confiável. Adicionalmente, a estatística do modelo sugere que este seja adequado para o estudo de tendências associadas com a variação de retornos de ativos e de mercados, bem como para avaliação de prêmios de risco baseados em dados históricos de tipo intervalo.

A segunda aborda regressão não-linear simétrica para tratar dados de natureza intervalar. As características desse modelo permitem que tanto a estimativa dos seus parâmetros quanto a predição de dados intervalares a partir do modelo ajustado, sejam menos sensíveis à presença de *outliers*. Também, o modelo SNLRM-IVD é flexível no que se refere ao tratamento dos erros do modelo, o que possibilita toda uma variedade de suposições probabilísticas e técnicas estatísticas para a descrição dos dados. Foram desenvolvidas simulações de Monte Carlo sob distribuição normal e t-Student considerando vários cenários com diferentes percentuais de *outliers* no conjunto de dados para verificar o comportamento do modelo em condições extremas. Em geral, as simulações realizadas indicam que o modelo proposto oferece uma melhora estatisticamente significativa na área de predição de dados simbólicos de tipo intervalo contaminados com dados aberrantes, em relação ao modelo poposto por LIMA NETO; CARVALHO (2016). Além das

simulações, a aplicação a um conjunto de dados reais foi feita sob a suposição de distribuição normal e t-Student (5 d.f.) para os erros. Nos modelos ajustados, os parâmetros estimados apresentaram valores de p menores que 0,05 em todos os casos, mas o valor da MMRE para SNLRM-IVD foi menor do que para NLM-IVD, o que também sugere que o tratamento dos erros com a distribuição t-Student para os erros é mais conveniente na presença de *outliers*, no âmbito de bases de dados simbólicas.

Ambas propostas usam um único modelo de regressão para ajustar os centros e amplitudes dos intervalos, melhorando a precisão do erro padrão das estimativas dos parâmetros e a avaliação foi baseada no critério da magnitude média dos erros relativos.

5.1 Contribuições

De forma explícita, as principais contribuições deste estudo foram:

- Introdução do modelo de precificação de ativos de capital para dados simbólicos de tipo intervalo.
- Interpretação do β intervalar no modelo *i*CAPM para avaliar as possibilidades dos investidores no mercado de ações.
- Ajuste dos centros e amplitudes dos intervalos através de um único modelo de regressão.
- Aplicação do modelo *i*CAPM proposto a dados simbólicos de tipo intervalo reais.
- Introdução do modelo de regressão não-linear simétrica para dados simbólicos de tipo intervalo.
- Possibilidade de construção de intervalos de confiança e teste de hipóteses sobre os parâmetros estimados do modelo SNLRM-IVD proposto.
- Aplicação do modelo SNLRM-IVD proposto a dados simbólicos de tipo intervalo reais.

5.2 Perspectivas para trabalhos futuros

1. Estender as técnicas propostas nesta dissertação para outros tipos de dados simbólicos.
2. Propor modelos de mistura de regressão simétrica usando funções lineares e não-lineares para conjuntos de dados simbólicos de tipo intervalo.

Referências

ANDERSON, T.; FANG, K.-T. Inference in multivariate elliptically contoured distributions based on maximum likelihood. **Statistical inference in elliptically contoured and related distributions**, [S.l.], p.201–216, 1990.

ARROYO, J.; MATÉ, C. Introducing interval time series: accuracy measures. **COMPSTAT 2006, proceedings in computational statistics**, [S.l.], p.1139–1146, 2006.

ARROYO, J.; MATÉ, C. Forecasting histogram time series with k-nearest neighbours methods. **International Journal of Forecasting**, [S.l.], v.25, n.1, p.192–207, 2009.

BARTHOLDY, J.; PEARE, P. Estimating cost of equity. **Available at SSRN 252270**, [S.l.], 2000.

BEATON, A. E.; TUKEY, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. **Technometrics**, [S.l.], v.16, n.2, p.147–185, 1974.

BELSON, W. A. Matching and prediction on the principle of biological classification. **Applied statistics**, [S.l.], p.65–75, 1959.

BERTRAND, P.; GOUPIL, F. Descriptive statistics for symbolic data. In: **Analysis of symbolic data**. [S.l.]: Springer, 2000. p.106–124.

BILLARD, L. Dependencies in bivariate interval-valued symbolic data. In: **Classification, Clustering, and Data Mining Applications**. [S.l.]: Springer, 2004. p.319–324.

BILLARD, L. Symbolic data analysis: what is it? In: **Compstat 2006-Proceedings in Computational Statistics**. [S.l.]: Springer, 2006. p.261–269.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: **Data Analysis, Classification, and Related Methods**. [S.l.]: Springer, 2000. p.369–374.

BILLARD, L.; DIDAY, E. Symbolic regression analysis. In: **Classification, Clustering, and Data Analysis**. [S.l.]: Springer, 2002. p.281–288.

BILLARD, L.; DIDAY, E. **Symbolic Data Analysis**: definitions and examples. 2003.

BILLARD, L.; DIDAY, E. Symbolic Data Analysis: conceptual statistics and data mining. **England: Wiley & Sons Ltd**, [S.l.], 2006.

BOCK, H. H. Automatische klassifikation. **Studia Mathematica/Mathematische Lehrbücher**, [S.l.], v.24, 1974.

BOCK, H.-H. Clustering algorithms and Kohonen maps for symbolic data. **Journal of the Japanese Society of Computational Statistics**, [S.l.], v.15, n.2, p.217–229, 2002.

BOCK, H.-H.; DIDAY, E. **Analysis of symbolic data**: exploratory methods for extracting statistical information from complex data. [S.l.]: Springer Science & Business Media, 2012.

CARVALHO, F. d. A. de et al. Clustering symbolic interval data based on a single adaptive Hausdorff distance. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS, 2007. **Anais...** [S.l.: s.n.], 2007. p.451–455.

CARVALHO, F. de; SOUZA, R. de. Unsupervised pattern recognition methods for interval data using non-quadratic distances. **Electronics Letters**, [S.l.], v.39, n.5, p.433–434, 2003.

CAZES, P. et al. Extension de l'analyse en composantes principales à des données de type intervalle. **Revue de Statistique appliquée**, [S.l.], v.45, n.3, p.5–24, 1997.

CHEUNG, S. Y.-L.; CHEUNG, Y.-W.; WAN, A. T. A high-low model of daily stock price ranges. **CESifo Working Paper Series**, [S.l.], n.2387, 2008.

CHMIELEWSKI, M. Elliptically symmetric distributions: a review and bibliography. **International Statistical Review/Revue Internationale de Statistique**, [S.l.], p.67–74, 1981.

CHOUAKRIA, A.; DIDAY, E.; CAZES, P. An improved factorial representation of symbolic objects. **Knowledge Extraction from Statistical Data**, [S.l.], v.301, p.305, 1998.

CIAMPI, A. et al. Growing a tree classifier with imprecise data. **Pattern Recognition Letters**, [S.l.], v.21, n.9, p.787–803, 2000.

COPELAND, T. E. et al. **Financial theory and corporate policy**. [S.l.]: Addison-Wesley Massachusetts, 1983. v.3.

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. [S.l.]: CRC Press, 1979.

CYSNEIROS, F. J. A. Local influence and residual analysis in heteroscedastic symmetrical linear models. In: STATISTICAL MODELLING: PROCEEDINGS OF THE 19TH INTERNATIONAL WORKSHOP ON STATISTICAL MODELLING, BIGGERI, A. **Anais...** [S.l.: s.n.], 2004. p.376–380.

CYSNEIROS, F. J. A.; PAULA, G. A. Restricted methods in symmetrical linear regression models. **Computational statistics & data analysis**, [S.l.], v.49, n.3, p.689–708, 2005.

DA SILVA, A. C. G. **Dissimilarity fuctions analysis based on dynamic clustering for symbolic data**. 2005.

DE CARVALHO, F. Histograms in symbolic data analysis. **Annals of Operations Research**, [S.l.], v.55, n.2, p.299–322, 1995.

DE CARVALHO, F. d. A. et al. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. **Pattern Recognition Letters**, [S.l.], v.27, n.3, p.167–179, 2006.

DIDAY, E. The symbolic approach in clustering and relating methods of data analysis: the basic choices. In: CONFERENCE OF THE INTERNATIONAL FEDERATION OF CLASSIFICATION SOCIETIES, 1. **Anais...** [S.l.: s.n.], 1987. p.673–684.

DIDAY, E. Introduction à l'approche symbolique en analyse des données. **Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle**, [S.l.], v.23, n.2, p.193–236, 1989.

DIDAY, E. Des objets de l'analyse des données à ceux de l'analyse des connaissances. **Induction Symbolique et Numérique à partir de données, Kodratoff Y. et Diday E. Eds., CEPADUES**, [S.l.], 1991.

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. **Wiley Interdisciplinary Reviews: Computational Statistics**, [S.l.], v.8, n.5, p.172–205, 2016.

DIDAY, E.; NOIRHOMME-FRAITURE, M. **Symbolic data analysis and the SODAS software**. [S.l.]: Wiley Online Library, 2008.

DIDAY, E.; SIMON, J. Clustering analysis. In: **Digital pattern recognition**. [S.l.]: Springer, 1980. p.47–94.

DOMINGUES, M. A. et al. Seleção de redes 4G: uma abordagem utilizando regressão simbólica simétrica. In: CONNEPI-CONGRESSO DE PESQUISA E INOVAçãO DA REDE NORTE NORDESTE, BELéM - PA. **Anais...** [S.l.: s.n.], 2009.

DOMINGUES, M. A.; SOUZA, R. M. de; CYSNEIROS, F. J. A. A symmetrical model applied to interval-valued data containing outliers with heavy-tail distribution. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING. **Anais...** [S.l.: s.n.], 2008. p.19–26.

DOMINGUES, M. A.; SOUZA, R. M. de; CYSNEIROS, F. J. A. A robust method for linear regression of symbolic interval data. **Pattern Recognition Letters**, [S.l.], v.31, n.13, p.1991–1996, 2010.

DOU, R.; ZONG, C.; LI, M. An interactive genetic algorithm with the interval arithmetic based on hesitation and its application to achieve customer collaborative product configuration design. **Applied Soft Computing**, [S.l.], v.38, p.384–394, 2016.

DOUZAL-CHOUAKRIA, A. **Extension des méthodes d'analyse factorielles à des données de type intervalle**. 1998. Tese (Doutorado em Ciência da Computação) — Paris IX Dauphine.

DUMONTIER, P. **Le modèle d'évaluation par arbitrage des actifs financiers**: une etude empirique sur le marché financier parisien. [S.l.]: Centre d'études et de recherches appliquées à la gestion, Université de Grenoble II, 1985.

EDWARDS, R. D.; MAGEE, J.; BASSETTI, W. **Technical analysis of stock trends**. [S.l.]: CRC Press, 2012.

ESTRADA, J. Systematic risk in emerging markets: the d-capm. **Emerging Markets Review**, [S.l.], v.3, n.4, p.365–379, 2002.

FAGUNDES, R. A.; DE SOUZA, R. M.; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. **Engineering Applications of Artificial Intelligence**, [S.l.], v.26, n.1, p.564–573, 2013.

FAMA, E. F. Efficient capital markets: ii. **The journal of finance**, [S.l.], v.46, n.5, p.1575–1617, 1991.

FAMA, E. F.; MACBETH, J. D. Risk, return, and equilibrium: empirical tests. **The journal of political economy**, [S.l.], p.607–636, 1973.

FAMÁ, R.; BARROS, L.; SILVEIRA, A. D. M. d. A Estrutura de Capital é Relevante? Novas Evidências a partir de dados norte-americanos e latino-americanos. **Caderno de Pesquisas em Administração**, [S.l.], v.8, n.2, p.71–84, 2001.

FANG, K.-T.; KOTZ, S.; NG, K. W. **Symmetric multivariate and related distributions**. [S.l.]: Chapman and Hall, 1990.

FRIEND, I.; WESTERFIELD, R.; GRANITO, M. New evidence on the capital asset pricing model. **The Journal of Finance**, [S.l.], v.33, n.3, p.903–917, 1978.

GALEA, M.; DÍAZ-GARCÍA, J. A.; VILCA, F. Influence diagnostics in the capital asset pricing model under elliptical distributions. **Journal of Applied Statistics**, [S.l.], v.35, n.2, p.179–192, 2008.

GALEA, M.; PAULA, G. A.; CYSNEIROS, F. J. A. On diagnostics in symmetrical nonlinear models. **Statistics & probability letters**, [S.l.], v.73, n.4, p.459–467, 2005.

GARCÍA-ASCANIO, C.; MATÉ, C. Electric power demand forecasting using interval time series: a comparison between var and imlp. **Energy Policy**, [S.l.], v.38, n.2, p.715–725, 2010.

GARMAN, M. B.; KLASS, M. J. On the estimation of security price volatilities from historical data. **Journal of business**, [S.l.], p.67–78, 1980.

GETTLER-SUMMA, M.; PARDOUX, C. Symbolic approaches for three-way data. In: **Analysis of Symbolic Data**. [S.l.]: Springer, 2000. p.342–354.

GIL, M. Á. et al. Testing linear independence in linear models with interval-valued data. **Computational Statistics & Data Analysis**, [S.l.], v.51, n.6, p.3002–3015, 2007.

GILBERT, S. Probabilités, analyse des données et statistique. **Paris, Éditions Technip**, [S.l.], 1990.

GOLDMAN, M. B.; SOSIN, H. B.; GATTO, M. A. Path dependent options:"buy at the low, sell at the high". **The Journal of Finance**, [S.l.], v.34, n.5, p.1111–1127, 1979.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. **Pattern Recognition**, [S.l.], v.24, n.6, p.567–578, 1991.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. **Pattern Recognition**, [S.l.], v.24, n.6, p.567–578, 1991.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new similarity measure. **IEEE Transactions on Systems, Man, and Cybernetics**, [S.l.], v.22, n.2, p.368–378, 1992.

GOWDA, K. C.; RAVI, T. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. **Pattern recognition**, [S.l.], v.28, n.8, p.1277–1282, 1995.

GURU, D.; KIRANAGI, B. B.; NAGABHUSHAN, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. **Pattern Recognition Letters**, [S.l.], v.25, n.10, p.1203–1213, 2004.

HICKEY, T.; JU, Q.; VAN EMDEN, M. H. Interval arithmetic: from principles to implementation. **Journal of the ACM (JACM)**, [S.l.], v.48, n.5, p.1038–1068, 2001.

ICHINO, M.; YAGUCHI, H. Generalized Minkowski metrics for mixed feature-type data analysis. **IEEE Transactions on Systems, Man, and Cybernetics**, [S.l.], v.24, n.4, p.698–708, 1994.

IRPINO, A. "Spaghetti" PCA analysis: an extension of principal components analysis to time dependent interval data. **Pattern recognition letters**, [S.l.], v.27, n.5, p.504–513, 2006.

JAGANNATHAN, R.; WANG, Z. The conditional CAPM and the cross-section of expected returns. **The Journal of finance**, [S.l.], v.51, n.1, p.3–53, 1996.

JENSEN, M. C.; BLACK, F.; SCHOLES, M. S. The capital asset pricing model: some empirical tests. In: **Studies in the Theory of Capital Markets**. [S.l.]: Praeger Publishers Inc, 1972.

LAURO, C. N.; PALUMBO, F. Principal component analysis of interval data: a symbolic data analysis approach. **Computational statistics**, [S.l.], v.15, n.1, p.73–87, 2000.

LAURO, N.; VERDE, R.; PALUMBO, F. Factorial discriminant analysis on symbolic objects. **Analysis of symbolic data Exploratory methods for extracting statistical information from complex data. Springer, Berlin Heidelberg New York**, [S.l.], p.212–233, 2000.

LEVY, H. The capital asset pricing model, inflation, and the investment horizon: the israeli experience. **Journal of Financial and Quantitative Analysis**, [S.l.], v.15, n.03, p.561–593, 1980.

LEWELLEN, J.; NAGEL, S. The conditional CAPM does not explain asset-pricing anomalies. **Journal of financial economics**, [S.l.], v.82, n.2, p.289–314, 2006.

LIMA NETO, E. d. A.; ANJOS, U. U. dos. Regression model for interval-valued variables based on copulas. **Journal of Applied Statistics**, [S.l.], v.42, n.9, p.2010–2029, 2015.

LIMA NETO, E. d. A.; CARVALHO, F. d. A. de. Centre and Range method for fitting a linear regression model to symbolic interval data. **Computational Statistics & Data Analysis**, [S.l.], v.52, n.3, p.1500–1515, 2008.

LIMA NETO, E. d. A.; CARVALHO, F. d. A. de. Constrained linear regression models for symbolic interval-valued variables. **Computational Statistics & Data Analysis**, [S.l.], v.54, n.2, p.333–347, 2010.

LIMA NETO, E. d. A.; CARVALHO, F. d. A. de. Nonlinear regression applied to interval-valued data. **Pattern Analysis and Applications**, [S.l.], p.1–16, 2016.

LIMA NETO, E. d. A.; CORDEIRO, G. M.; CARVALHO, F. d. A. de. Bivariate symbolic regression models for interval-valued variables. **Journal of Statistical Computation and Simulation**, [S.l.], v.81, n.11, p.1727–1744, 2011.

LINTNER, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. **The review of economics and statistics**, [S.l.], p.13–37, 1965.

LLATAS, M. B.; M., G.-S. J. Segmentation trees for stratified data. In: **Analysis of Symbolic Data**. [S.l.]: Springer, 2000. p.266–293.

LUGER, G. F. **Artificial intelligence**: structures and strategies for complex problem solving. [S.l.]: Pearson education, 2005.

MAIA, A. L. S.; CARVALHO, F. d. A. de. Fitting a least absolute deviation regression model on interval-valued data. In: BRAZILIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE. **Anais...** [S.l.: s.n.], 2008. p.207–216.

MAIA, A. L. S.; CARVALHO, F. d. A. de; LUDERMIR, T. B. A hybrid model for symbolic interval time series forecasting. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING. **Anais...** [S.l.: s.n.], 2006. p.934–941.

MAIA, A. L. S.; CARVALHO, F. d. A. de; LUDERMIR, T. B. Forecasting models for interval-valued time series. **Neurocomputing**, [S.l.], v.71, n.16, p.3344–3352, 2008.

MAIOR, V. Q. S.; CYSNEIROS, F. J. A. Estimação do risco sistemático em modelos CAPM com erros normais assimétricos. **Rev. Bras. Biom**, [S.l.], v.27, n.2, p.197–209, 2009.

MALKIEL, B. G.; FAMA, E. F. Efficient capital markets: a review of theory and empirical work. **The journal of Finance**, [S.l.], v.25, n.2, p.383–417, 1970.

MARKOWITZ, H. Portfolio selection. **The journal of finance**, [S.l.], v.7, n.1, p.77–91, 1952.

MBALLO, C.; DIDAY, E. Decision trees on interval valued variables. **The electronic journal of symbolic data analysis**, [S.l.], v.3, n.1, p.8–18, 2005.

MERTON, R. C. An intertemporal capital asset pricing model. **Econometrica: Journal of the Econometric Society**, [S.l.], p.867–887, 1973.

MERTON, R. C. On estimating the expected return on the market: an exploratory investigation. **Journal of financial economics**, [S.l.], v.8, n.4, p.323–361, 1980.

MICHALSKI, R. S.; STEPP, R. E.; DIDAY, E. **A recent advance in data analysis**: clustering objects into classes characterized by conjunctive concepts. [S.l.: s.n.], 1981.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2015.

MOORE, R. E. **Interval arithmetic and automatic error analysis in digital computing**. [S.l.]: DTIC Document, 1962.

MOORE, R. E.; KEARFOTT, R. B.; CLOUD, M. J. **Introduction to interval analysis**. [S.l.]: Siam, 2009.

MORGAN, J. N.; SONQUIST, J. A. Problems in the analysis of survey data, and a proposal. **Journal of the American statistical association**, [S.l.], v.58, n.302, p.415–434, 1963.

MORINEAU, A. et al. Analyses des données et modélisation des séries temporelles. Application à la prévision des ventes de périodiques. **Revue de statistique appliquée**, [S.l.], v.42, n.4, p.61–81, 1994.

MOSSIN, J. Equilibrium in a capital asset market. **Econometrica: Journal of the econometric society**, [S.l.], p.768–783, 1966.

MUKHERJI, S. The capital asset pricing model's risk-free rate. **The International Journal of Business and Finance Research**, [S.l.], v.5, n.2, p.75–83, 2011.

NAGABHUSHAN, P.; GOWDA, K. C.; DIDAY, E. Dimensionality reduction of symbolic data. **Pattern recognition letters**, [S.l.], v.16, n.2, p.219–223, 1995.

NAKAMURA, W. T. **Eficiência da carteira teórica do índice Bovespa no contexto da moderna teoria de carteiras**. 1998. Tese (Doutorado em Ciência da Computação) — .

NOIRHOMME-FRAITURE, M.; BRITO, P. Far beyond the classical data models: symbolic data analysis. **Statistical Analysis and Data Mining**, [S.l.], v.4, n.2, p.157–170, 2011.

NOIRHOMME-FRAITURE, M.; ROUARD, M. Zoom Star: a solution to complex statistical object representation. In: HUMAN-COMPUTER INTERACTION INTERACT'97. **Anais...** [S.l.: s.n.], 1997. p.100–101.

OLIVEIRA, P. W.; DIVERIO, T. A.; CLAUDIO, D. M. **Fundamentos da matemática intervalar**. [S.l.]: Sagra-Luzzatto, 1997.

PARKINSON, M. The extreme value method for estimating the variance of the rate of return. **Journal of Business**, [S.l.], p.61–65, 1980.

PAULA, G. A.; CYSNEIROS, F. J. A. Systematic risk estimation in symmetric models. **Applied Economics Letters**, [S.l.], v.16, n.2, p.217–221, 2009.

PIAMSUWANNAKIT, S. et al. Capital Asset Pricing Model with Interval Data. In: INTERNATIONAL SYMPOSIUM ON INTEGRATED UNCERTAINTY IN KNOWLEDGE MODELLING AND DECISION MAKING. **Anais...** [S.l.: s.n.], 2015. p.163–170.

RIBENBOIM, G. Teste de modelo CAPM com dados brasileiros. **Finanças aplicadas ao Brasil**, [S.l.], v.2, 2002.

ROLL, R. A critique of the asset pricing theory's tests Part I: on past and potential testability of the theory. **Journal of financial economics**, [S.l.], v.4, n.2, p.129–176, 1977.

RUPPERT, D. **Statistics and finance**: an introduction. [S.l.]: Springer Science & Business Media, 2004.

RUSSELL, S. J. et al. **Artificial intelligence**: a modern approach. [S.l.]: Prentice hall Upper Saddle River, 2003. v.2.

SHARPE, W. F. A simplified model for portfolio analysis. **Management science**, [S.l.], v.9, n.2, p.277–293, 1963.

SHARPE, W. F. Capital asset prices: a theory of market equilibrium under conditions of risk. **The journal of finance**, [S.l.], v.19, n.3, p.425–442, 1964.

SILVA, C. A. T.; MUNHOZ, D. A. A utilização do lucro contábil como proxy do risco no Brasil. **Encontro da Associação Nacional de Pós-graduação e Pesquisa em Administração**, [S.l.], v.30, 2006.

SNEATH, P. H.; SOKAL, R. R. Numerical taxonomy. **Nature**, [S.l.], v.193, n.4818, p.855–860, 1962.

SOUZA, R. M. de; DE CARVALHO, F. d. A. Clustering of interval data based on city–block distances. **Pattern Recognition Letters**, [S.l.], v.25, n.3, p.353–365, 2004.

SOUZA, R. M. de; QUEIROZ, D. C.; CYSNEIROS, F. J. A. Logistic regression-based pattern classifiers for symbolic interval data. **Pattern Analysis and Applications**, [S.l.], v.14, n.3, p.273–282, 2011.

TEAM, R. C. **R**: a language and environment for statistical computing. r foundation for statistical computing, vienna, austria. 2013. [S.l.]: ISBN 3-900051-07-0, 2014.

VANEGAS, L. H.; CYSNEIROS, F. J. A. Assessment of diagnostic procedures in symmetrical nonlinear regression models. **Computational Statistics & Data Analysis**, [S.l.], v.54, n.4, p.1002–1016, 2010.

XU, W. **Symbolic data analysis**: interval-valued data regression. 2010. Tese (Doutorado em Ciência da Computação) — PhD thesis, University of Georgia.

A

Apéndice A. Figuras do Capítulo 4

A seguir, são apresentados os gráficos dos dados simulados nos experimentos, realzados para o estudo do capítulo 4, relativo aos cenários I e II com 1% e 5% de *outliers*.

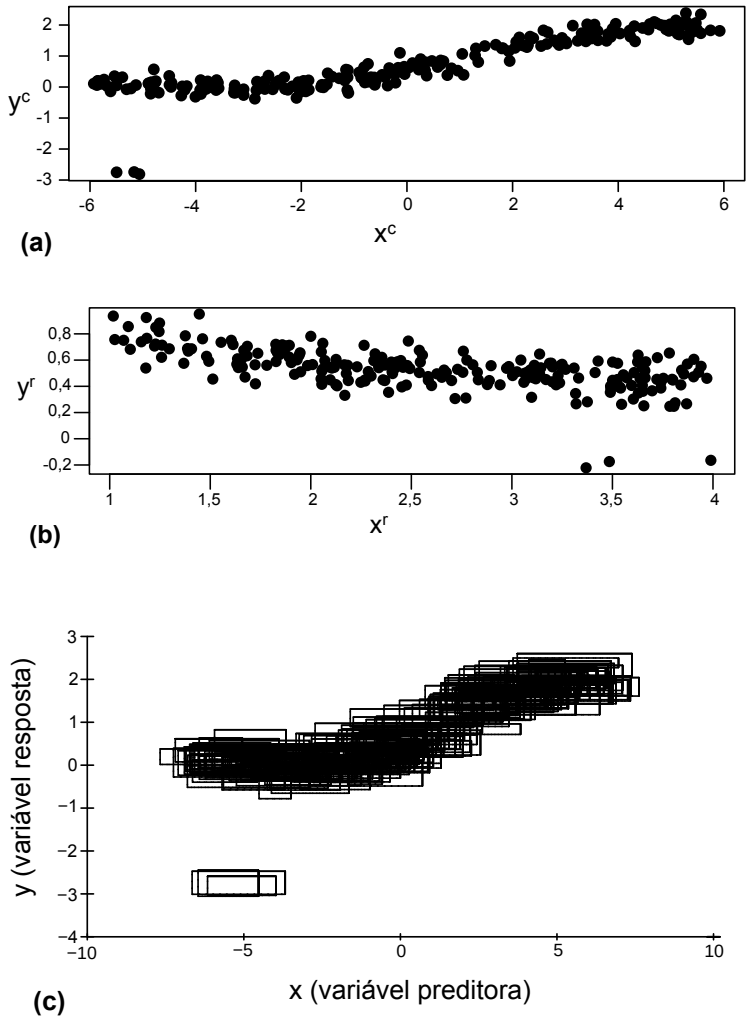

Figura A.1: Conjunto de dados simulados para a Configuração 1, cenário I, com 1% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 1% de elementos inferiores como *outliers*, (b) Ranges dos intervalos com 1% de elementos superiores como *outliers*, (c) Intervalos.

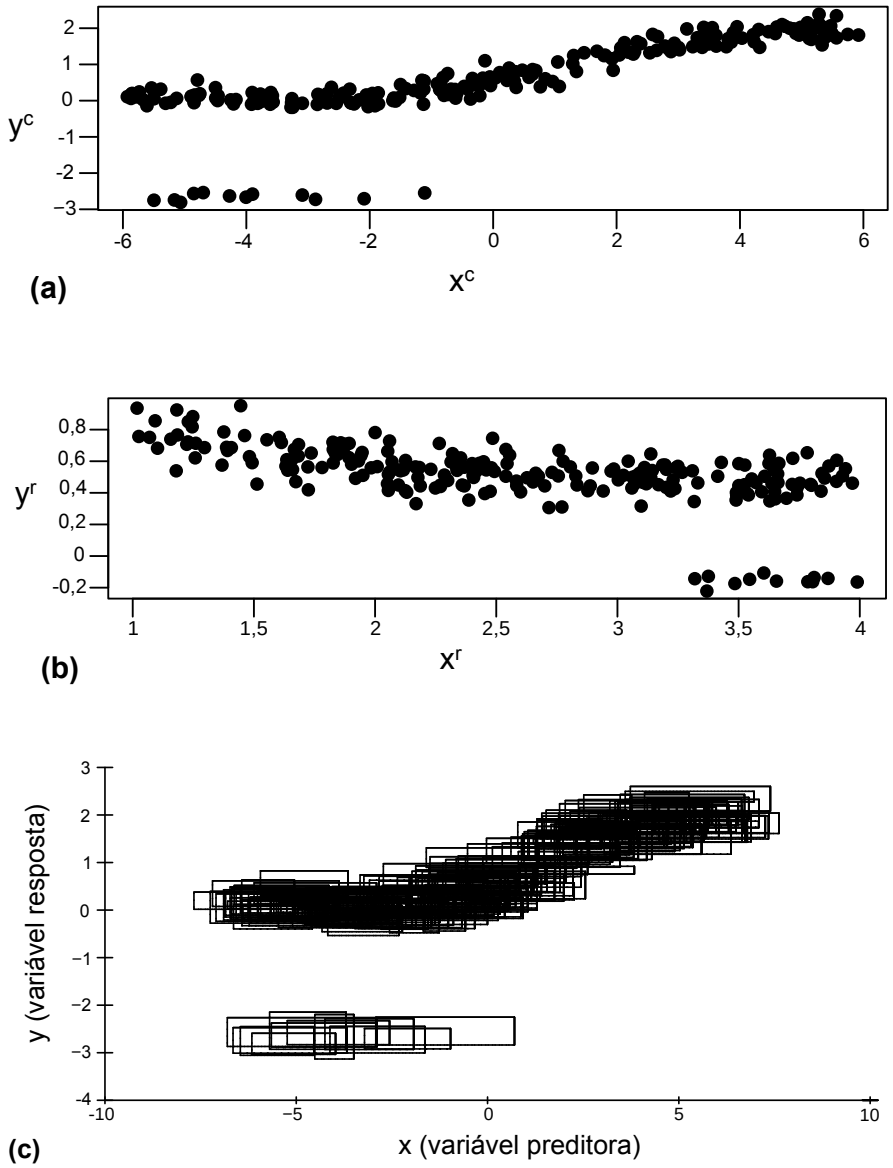

Figura A.2: Conjunto de dados simulados para a Configuração 1, cenário I, com 5% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 5% de elementos inferiores como *outliers*, (b) Ranges dos intervalos com 5% de elementos inferiores como *outliers*, (c) Intervalos.

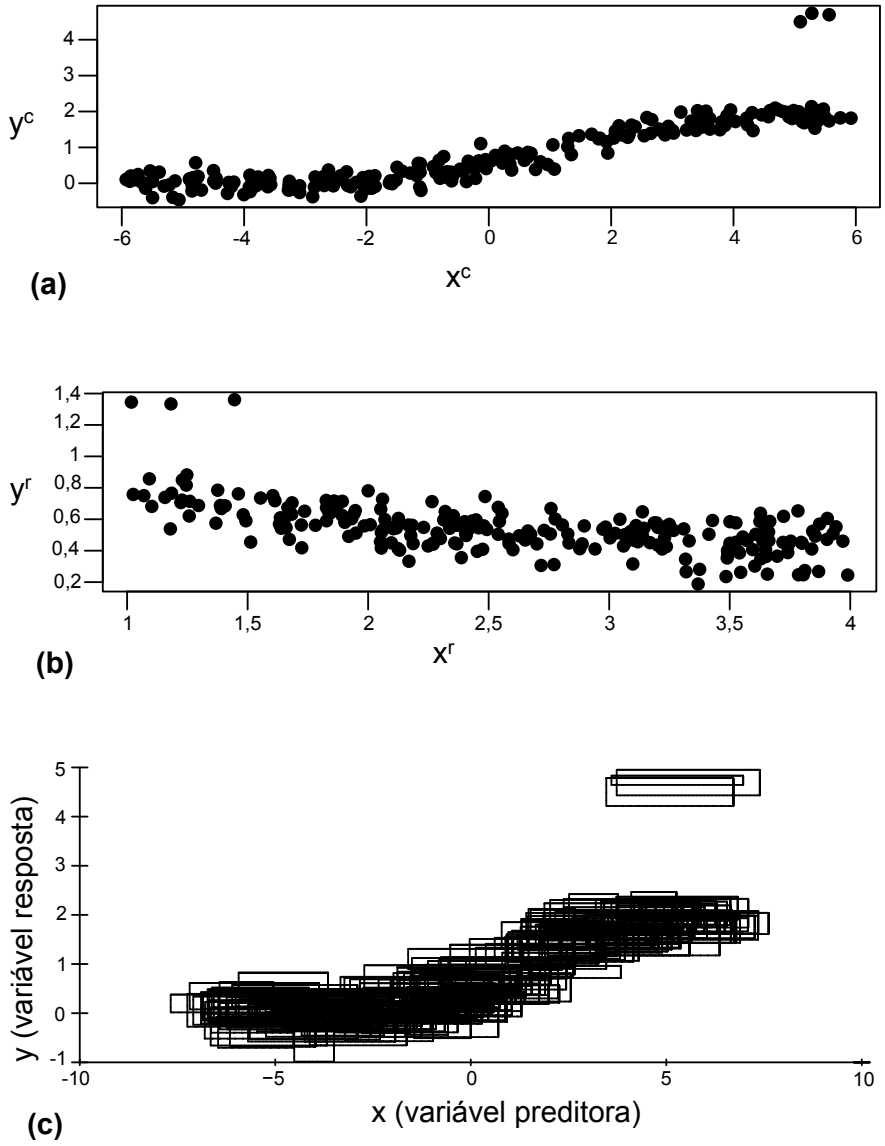

Figura A.3: Conjunto de dados simulados para a Configuração 1, cenário II, com 1% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 1% de elementos superiores como *outliers*, (b) Ranges dos intervalos com 1% de elementos superiores como *outliers*, (c) Intervalos.

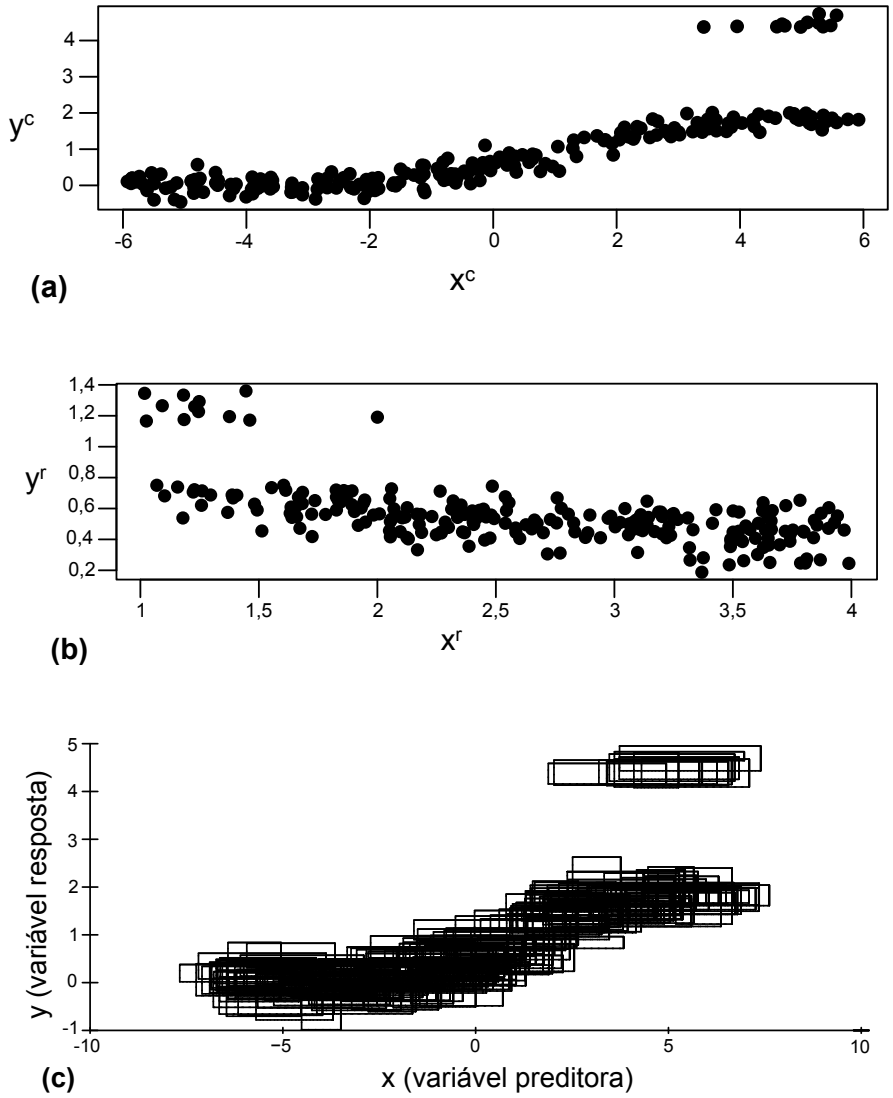

Figura A.4: Conjunto de dados simulados para a Configuração 1, cenário II, com 5% de *outliers* intervalares. Os gráficos mostram a: (a) Centros dos intervalos com 5% de elementos superiores como *outliers*, (b) Ranges dos intervalos com 5% de elementos superiores como *outliers*, (c) Intervalos.

B

Apéndice B. Implementação em R

A seguir, são apresentadas as principais sub-rotinas implementadas em linguagem R para os métodos propostos nos Capítulos 3 e 4 desta dissertação. Cabe ressaltar que o programa completo pode ser solicitado aos autores e, além das rotinas, estão disponíveis os complementos seguintes:

- A livraria *Elliptical*.
- O segundo exemplo mostrado no Capítulo 3.
- A geração de todos os cenários mostrados no Capítulo 4.
- A geração dos plots intervalares.
- Implementação dos testes de hipóteses para amostras emparelhadas.
- Entre outras rotinas.

*i*CAPM PARA DADOS DA MICROSOFT

```
library (ISDA.R)
intervalar = read.table ("Microsoft.csv", header = T, sep=";", as.is=T)
View (intervalar)

# calculo dos retornos
yt_inf <- NULL
yt_sup <- NULL
i <- nrow(intervalar)
while (i>1)
{
yt_inf <- ((intervalar$Low_Microsoft [i] - intervalar$High_Mcrosoft [i-1])/intervalar$High_Mcrosoft
[i-1])*100
```

```
yt_sup <- ((intervalar$High_Mcrosoft [i]- intervalar$Low_Microsoft [i-1])/intervalar$Low_Microsoft
[i-1])*100
intervalar$yt_Inf[i] <- yt_inf
intervalar$yt_Sup [i] <- yt_sup
i = i-1
}
rmt_inf <- NULL
rmt_sup <- NULL
i <- nrow(intervalar)
while (i>1)
{
rmt_inf <- ((intervalar$Low_S.P.500 [i] - intervalar$High_S.P.500[i-1])/intervalar$High_S.P.500[i-
1])*100
rmt_sup <- ((intervalar$High_S.P.500 [i]- intervalar$Low_S.P.500 [i-1])/intervalar$Low_S.P.500
[i-1])*100
intervalar$rmt_Inf[i] <- rmt_inf
intervalar$rmt_Sup [i] <- rmt_sup
i = i-1
}
intervalar <- intervalar [-c(1),]
intervalar$rft_Inf <- intervalar$Low_T.Bill*100
intervalar$rft_Sup <- intervalar$High_T.Bill*100
View (intervalar)

# calculo do prêmio de risco
Prmt_inf <- NULL
Prmt_sup <- NULL
for (i in 1:nrow(intervalar))
{
Prmt_inf <- intervalar$rmt_Inf[i] - intervalar$rft_Sup[i]
Prmt_sup <- intervalar$rmt_Sup[i] - intervalar$rft_Inf[i]
intervalar$Prmt_Inf[i] <- Prmt_inf
intervalar$Prmt_Sup[i] <- Prmt_sup
}
Pyt_inf <- NULL
Pyt_sup <- NULL
for (i in 1:nrow(intervalar))
{
Pyt_inf <- intervalar$yt_Inf[i] - intervalar$rft_Sup[i]
```

```
Pyt_sup <- intervalar$yt_Sup[i] - intervalar$rft_Inf[i]
intervalar$Pyt_Inf[i] <- Pyt_inf
intervalar$Pyt_Sup[i] <- Pyt_sup
}
# Regresão centro e range
intervalar$Premio_yt_Centro <- (intervalar$Pyt_Inf + intervalar$Pyt_Sup)/2
intervalar$Premio_rmt_Centro <- (intervalar$Prmt_Inf + intervalar$Prmt_Sup)/2
intervalar$Premio_rmt_range <- (intervalar$Prmt_Sup - intervalar$Prmt_Inf)/2
intervalar$Premio_yt_range <- (intervalar$Pyt_Sup - intervalar$Pyt_Inf)/2
y <- c(intervalar$Premio_yt_Centro, intervalar$Premio_yt_range)
nc <- length(intervalar$Premio_yt_Centro)
nr <- length(intervalar$Premio_yt_range)
x1 <- c(rep(1,nc),rep(0,nr))
x2 <- c(rep(0,nc),rep(1,nr))
x3 <- c(intervalar$Premio_rmt_Centro,rep(0,nr))
x4 <- c(rep(0,nc),intervalar$Premio_rmt_range)
X <- cbind(x2, x3, x4)
regresion <- lm (y X-1)
summary (regresion)
```

CONFIGURAÇÃO 2 e SNLRM

```
library(Matrix)
library (ISDA.R)

#Matriz de Derivadas
DerSim1 <- function(parm,X)
{
alpha <- parm[1]
beta <- parm[2]
alpha1 <- parm[3]
beta1 <- parm[4]
x0 <- X[,1]
x <- X[,2]
x1 <- X[,3]
x2 <- X[,4]
.grad <- array(0, c(length(x0), 4), list(NULL, c("alpha","beta","alpha1","beta1")))
.grad[, "alpha"] <- -(x1*beta*x)/(alpha+x)2̂
.grad[, "beta"] <- (x1*x)/(alpha + x)
```

```
.grad[, "alpha1"] <- -(x2*beta1*x)/(alpha1+x)^2
.grad[, "beta1"] <- (x2*x)/(alpha1 + x)
.value <- x1*(beta*x)/(x + alpha) + x2*(beta1*x)/(x + alpha1)
.hess <- array(0, c(4, 4,length(x0)),
list( c("alpha","beta","alpha1","beta1"), c("alpha","beta","alpha1","beta1"),NULL))
.hess["alpha","alpha", ]<- (2*x1*beta*x)/(alpha+x)^3
.hess["alpha","beta", ] <- -(x1*x)/(alpha+x)^2
.hess["alpha","alpha1", ]<- rep(0,length(x0))
.hess["alpha","beta1", ]<- rep(0,length(x0))
.hess["beta","alpha", ] <- .hess["alpha","beta", ]
.hess["beta","beta", ] <- rep(0,length(x0))
.hess["beta","alpha1", ] <- rep(0,length(x0))
.hess["beta","beta1", ] <- rep(0,length(x0))
.hess["alpha1","alpha", ] <- .hess["alpha","alpha1", ]
.hess["alpha1","beta", ] <- .hess["beta","alpha1", ]
.hess["alpha1","alpha1", ] <- (2*x2*beta1*x)/(alpha1+x)^3
.hess["alpha1","beta1", ] <- -(x2*x)/(alpha1+x)^2
.hess["beta1","alpha", ] <- .hess["alpha","beta1", ]
.hess["beta1","beta", ] <- .hess["beta","beta1", ]
.hess["beta1","alpha1", ] <- .hess["alpha1","beta1", ]
.hess["beta1","beta1", ] <- rep(0,length(x0))
fit <- list(value=.value,gradient=.grad, hessian=.hess)
fit
}

n <- 300
N <- 75
sigmac <- 5
sigmar <- 5
MMREN <- numeric(1000)
MMRET <- numeric(1000)
result <- rep(100000,8)
mc <- 0
for (j in 1:50)
{
if (mc == 0) {
alphaCentro <- runif(1, 19, 21)
betaCentro <- runif(1, 117, 123)
alphaRange <- runif(1, 19, 21)
```

```
betaRange <- runif(1, 117, 123)
}
for (i in 1:1000)
{
erroCentro <- rnorm(n)
erroRange <- rnorm(n)
xCentro <- runif(n, 5, 210)
xRange <- runif(n, 5, 210)
yCentro <- (betaCentro*xCentro)/(xCentro + alphaCentro) + sqrt(sigmac)*erroCentro
yRange <- (betaRange*xRange)/(xRange + alphaRange) + sqrt(sigmac)*erroRange
simulacoes <- data.frame(yCentro,yRange,xCentro,xRange)

# particionamento aleatorio de dados
dim(simulacoes)
#Sample Indexes
Indices = sample(1:nrow(simulacoes), size=0.75*nrow(simulacoes))
# Split data
train <- simulacoes[Indices,]
test <- simulacoes[-Indices,]

YcentroN <- train[,1]
YrangeN <- train[,2]
xcentroG <- train[,3]
xrangeG <- train[,4]

# OUTLIERS
IndicesOutliers = sample(1:nrow(train), size=22)
for(k in 1:22)
{
if(k <= 11){
YcentroN[IndicesOutliers[k]] <- YcentroN[IndicesOutliers[k]] + (3 * sd(YcentroN))
YrangeN[IndicesOutliers[k]] <- YrangeN[IndicesOutliers[k]] + (3 * sd(YrangeN))
}
if(k > 11){
YcentroN[IndicesOutliers[k]] <- YcentroN[IndicesOutliers[k]] - (3 * sd(YcentroN))
YrangeN[IndicesOutliers[k]] <- YrangeN[IndicesOutliers[k]] - (3 * sd(YrangeN))
}
}
```

```
nc1 <- length(YcentroN)
nr1 <- length(YrangeN)
X1 <- c(rep(1,nc1),rep(0,nr1))
X2 <- c(rep(0,nc1),rep(1,nr1))
Xn <- c(xcentroG,xrangeG)
XGeral <- cbind(Xn, X1, X2)

#Ajuste do modelo com erro Student
fitT1 <- elliptical(formula=y1_normal1 XGeral,linear=F, DerB=DerSim1,
parmB=c(alphaCentro, betaCentro, alphaRange, betaRange),
family=Student(4),data=dados1_normal1)

#predição com teste
coefT1 <- fitT1$coefficients
y_predita_CT1 <- (coefT1[2]*test[,3])/(test[,3] + coefT1[1])
y_predita_RT1 <- (coefT1[4]*test[,4])/(test[,4] + coefT1[3])

#intervalos para y teste
Y_inf = test[,1] - test[,2]/2
Y_sup = test[,1] + test[,2]/2

# Intervalos preditos para y teste
YTpred1_inf1 = y_predita_CT1 - y_predita_RT1/2
YTpred1_sup1 = y_predita_CT1 + y_predita_RT1/2

#MMRE Student
mmreT <- (1/(2*N))*sum((abs((Y_inf-YTpred1_inf1)/Y_inf) + abs((Y_sup-YTpred1_sup1)/Y_sup)))
MMRET[i] <- mmreT
}
if (mean(MMRET)<result[1]){
result[1] <- mean(MMRET)
result[3] <- sd(MMRET)
result[4] <- alphaCentro
result[5] <- betaCentro
result[6] <- alphaRange
result[7] <- betaRange
}
else
{
```

```
mc <- 0
}
}
result
```

SNLRM COM DADOS REAIS

```
library(Matrix)
library (ISDA.R)

# Dados
medical = read.table ("Medical.csv", header = T, sep=";", as.is=T)
y_inf<-medical$y_min
y_sup<-medical$y_max
x_inf<-medical$x_min
x_sup<-medical$x_max

yCentro <- (y_sup + y_inf)/2
yRange <- (y_sup - y_inf)/2
xCentro <- (x_sup + x_inf)/2
xRange <- (x_sup - x_inf)/2

#Matriz de Derivadas
DerSim <- function(parm,XGeral)
{
alpha <- parm[1]
beta <- parm[2]
alpha1 <- parm[3]
beta1 <- parm[4]
x0 <- XGeral[,1]
x <- XGeral[,2]
x1 <- XGeral[,3]
x2 <- XGeral[,4]
.grad <- array(0, c(length(x0), 4), list(NULL, c("alpha","beta","alpha1","beta1")))
.grad[, "alpha"] <- x1*(1 - exp(-beta*x))
.grad[, "beta"] <- x1*alpha*x*exp(-beta*x)
.grad[, "alpha1"] <- x2*(x2̂)
.grad[, "beta1"] <- x2/x
.value <- x1*(alpha*(1-exp(-beta*x))) + x2*(alpha1*(x2̂))+(beta/x)
```

```
.hess <- array(0, c(4, 4,length(x0)),
list( c("alpha","beta","alpha1","beta1"), c("alpha","beta","alpha1","beta1"),NULL))
.hess["alpha","alpha", ]<- rep(0,length(x0))
.hess["alpha","beta", ] <- x1*x*exp(-beta*x)
.hess["alpha","alpha1", ]<- rep(0,length(x0))
.hess["alpha","beta1", ]<- rep(0,length(x0))
.hess["beta","alpha", ] <- .hess["alpha","beta", ]
.hess["beta","beta", ] <- -alpha*(x2̂)*x1*exp(-beta*x)
.hess["beta","alpha1", ] <- rep(0,length(x0))
.hess["beta","beta1", ] <- rep(0,length(x0))
.hess["alpha1","alpha", ] <- .hess["alpha","alpha1", ]
.hess["alpha1","beta", ] <- .hess["beta","alpha1", ]
.hess["alpha1","alpha1", ] <- rep(0,length(x0))
.hess["alpha1","beta1", ] <- rep(0,length(x0))
.hess["beta1","alpha", ] <- .hess["alpha","beta1", ]
.hess["beta1","beta", ] <- .hess["beta","beta1", ]
.hess["beta1","alpha1", ] <- .hess["alpha1","beta1", ]
.hess["beta1","beta1", ] <- rep(0,length(x0))
fit <- list(value=.value,gradient=.grad, hessian=.hess)
fit
}

#Leave One Out
i <- NULL
y_predita_CN <- NULL
y_predita_CT <- NULL
y_predita_RN <- NULL
y_predita_RT <- NULL
mmreT <- NULL
mmreN <- NULL
for(i in 1:42)
{
# Split data
train <- medical[-i,]
test <- medical[i,]
y_inf<-train$y_min
y_sup<-train$y_max
x_inf<-train$x_min
x_sup<-train$x_max
```

```
yCentro <- (y_sup + y_inf)/2
yRange <- (y_sup - y_inf)/2
xCentro <- (x_sup + x_inf)/2
xRange <- (x_sup - x_inf)/2

xCteste <- (test[,4] + test[,3])/2
xRteste <- (test[,4] - test[,3])/2

Y <- c(yCentro, yRange)
nc <- length(yCentro)
nr <- length(yRange)
X1 <- c(rep(1,nc),rep(0,nr))
X2 <- c(rep(0,nc),rep(1,nr))
Xn <- c(xCentro,xRange)
XGeral <- cbind(Xn, X1, X2) geral <- data.frame(Y,XGeral)

#Ajuste do modelo com erro Student
fit_T <- elliptical(formula=Y XGeral, linear=F, DerB=DerSim,
parmB=c(121.74379, 0.16317, 0.1840, 49.8575),
family=Student(5),data=geral)
coef_T <- fit_T$coefficients

#predição y_predita_CT <- coef_T[1] * (1 - exp(-coef_T[2]*xCteste))
y_predita_RT <- coef_T[3]* (xRteste)2̂ + (coef_T[4]/(xRteste))
y_pred_inf_T <- y_predita_CT - y_predita_RT/2
y_pred_sup_T <- y_predita_CT + y_predita_RT/2

#MMRE Student
mmreT[i] <- (abs((y_inf[i]-y_pred_inf_T)/y_inf[i])) + (abs((y_sup[i]-y_pred_sup_T)/y_sup[i]))
}
MMRET <- (1/82)*sum(mmreT[1:41])
summary(fit_T)
```