Pontifícia Universidade Católica
do Rio de Janeiro

André Luiz Farias Novaes

Programação Genética Econométrica: Uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientador: Prof. Ricardo Tanscheit
Co-orientador: Dr. Douglas Mota Dias

Rio de Janeiro
Abril de 2015

Pontifícia Universidade Católica
do Rio de Janeiro

André Luiz Farias Novaes

Programação Genética Econométrica: Uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ricardo Tanscheit
Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Dr. Douglas Mota Dias
Co-orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. Cristiano Augusto Coelho Fernandes
Departamento de Engenharia Elétrica – PUC-Rio

Prof. André Vargas Abs da Cruz
UEZO

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico

Rio de Janeiro, 13 de abril de 2015

André Luiz Farias Novaes

Graduou-se em Engenharia de Produção pela Pontifícia Universidade Católica do Rio de Janeiro em 2010. Possui especialização em finanças pela COPPEAD/UFRJ.

Ficha Catalográfica

Novaes, André Luiz Farias

Programação genética econométrica: uma nova abordagem para problemas de regressão e classificação em conjuntos de dados seccionais / André Luiz Farias Novaes; orientador: Ricardo Tanscheit; co-orientador: Douglas Mota Dias. – 2015.

125 f. : il. (color.) ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2015.

Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Programação genética. 3. Econometria em dados seccionais. 4. Regressão e classificação. I. Tanscheit, Ricardo. II. Dias, Douglas Mota. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

A todos que de alguma forma fizeram parte desta jornada.

Agradecimentos

Muitos participaram da autoria deste trabalho. Embora esta seja uma seção de agradecimentos, o mais adequado seria nomeá-la de seção de “coautoria”, pois não há trabalho que seja feito sozinho.

Inicialmente, agradeço a Deus e ao mestre Jesus pela oportunidade de colaborar com um trabalho que, desejo muito, possa agregar valor incondicionalmente à vida das pessoas.

Às amigas e mestras Ana e Maria, e a todos os amigos e mestres que elas também representam, por terem me apresentado e convidado a seguir o caminho da evolução, corretude e justiça.

Aos orientadores Ricardo Tanscheit e Douglas Mota, pelo exemplo profissional, seriedade no trabalho desenvolvido e valiosas orientações na construção desta dissertação.

Ao professor Cristiano Fernandes, pela orientação e notoriedade com que leciona.

Ao professor André Vargas, pelo interesse em agregar valor ao trabalho realizado nesta dissertação.

À professora Marley Vellasco e Dom Pedrito, pela oportunidade de ingresso no DEE da PUC-Rio.

A todos os professores do DEE da PUC-Rio.

À Alcina, pelo profissionalismo e conduta no DEE da PUC-Rio.

Ao CNPq, pelo estímulo à pesquisa e financiamento.

Aos meus pais pelo grande exemplo, ensinando-me um grande caminho para construir valor: o ensino.

Ao meu grande amor Gabi, por seguir ao meu lado firmemente ao longo de todo o curso, e à Ana e Renato que, além das pessoas que são, terem me dado a oportunidade de viver ao lado dela.

Ao meu irmão Henrique, pelos momentos de descontração que tenho o privilégio de ter quando estou ao seu lado, e Carol, pelos mesmos motivos.

Aos que já partiram, em particular aos meus avôs e avós, que contribuíram de forma fundamental à minha formação como cidadão.

Ao tio Paulo, Patrícia e Sofia, por todo o apoio e momentos de descontração.

A todos os amigos e irmãos do grupo GEUDADE, pela companhia e instruções ao longo da caminhada.

Ao amigo Adriano Koshiyama, pelo exemplo de aluno e pesquisador com o qual tive o privilégio de conviver e aprender nos anos de mestrado.

Ao amigo Mauricio Cabrera, pelos sábios e sinceros conselhos em econometria.

Ao amigo Ricardo Vela, pela ajuda na busca pela verdade.

Aos amigos do Clube América, em especial a Pedro Nuno e Demétrius Medrado, por terem sido meus melhores amigos no dia a dia dessa caminhada.

Aos amigos Bruno Fânzeres e Alexandre Moreira.

Ao amigo Bruno Agrélio, pela ajuda teórica fundamental à dissertação.

Ao amigo Hugo Baldioti, pela grande ajuda na construção deste documento.

Aos amigos Alexandre Figueira, Abel Arrieta, Felipe Azevedo e Olivério Fernandes, pela ajuda ao longo do curso.

Aos amigos do DI da PUC-Rio.

Agradeço de coração a todos, considerando-os coautores desta dissertação.

Resumo

Novaes, André Luiz Farias; Tanscheit, Ricardo (Orientador); Dias, Douglas Mota (Co-Orientador). **Programação Genética Econométrica: Uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais.** Rio de Janeiro, 2015. 125p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Esta dissertação propõe modelos parcimoniosos para tarefas de regressão e classificação em conjuntos de dados exclusivamente seccionais, mantendo-se a hipótese de amostragem aleatória. Os modelos de regressão são lineares, estimados por Mínimos Quadrados Ordinários resolvidos pela Decomposição QR, apresentando solução única sob posto cheio ou não da matriz de regressores. Os modelos de classificação são não lineares, estimados por Máxima Verossimilhança utilizando uma variante do Método de Newton, nem sempre apresentando solução única. A parcimônia dos modelos de regressão é fundamentada na prova matemática de que somente agregará acurácia ao modelo o regressor que apresentar módulo da estatística de teste, em um teste de hipótese bicaudal, superior à unidade. A parcimônia dos modelos de classificação é fundamentada em significância estatística e embasada intuitivamente no resultado teórico da existência de classificadores perfeitos. A Programação Genética (PG) realiza o processo de evolução de modelos, explorando o espaço de busca de possíveis modelos, constituídos de distintos regressores. Os resultados obtidos via Programação Genética Econométrica (PGE) – nome dado ao algoritmo gerador de modelos – foram comparados aos proporcionados por *benchmarks* em oito distintos conjuntos de dados, mostrando-se competitivos em termos de acurácia na maior parte dos casos. Tanto sob o domínio da PG quanto sob o domínio da econometria, a PGE mostrou benefícios, como o auxílio na identificação de *introns*, o combate ao *bloat* por significância estatística e a geração de modelos econométricos de elevada acurácia, entre outros.

Palavras-chave

Programação Genética; Econometria em dados seccionais; Regressão e Classificação.

Abstract

Novaes, André Luiz Farias; Tanscheit, Ricardo (Advisor); Dias, Douglas Mota (Co-Advisor). **Econometric Genetic Programming: a New Approach for Regression and Classification Problems in Cross-Sectional Datasets.** Rio de Janeiro, 2015. 125p. MSc. Dissertation – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

This dissertation proposes parsimonious models for regression and classification tasks in cross-sectional datasets under random sample hypothesis. Regression models are linear in parameters, estimated by Ordinary Least Squares solved by QR Decomposition, presenting a unique solution under full rank of the regressor matrix or not. Classification models are nonlinear in parameters, estimated by Maximum Likelihood, not always presenting a unique solution. Parsimony in regression models is based on the mathematical proof that accuracy will be added to models only by the regressor that presents a test statistic module higher than a predefined value in a two-sided hypothesis test. Parsimony in classification models is based on statistical significance and, intuitively, on the theoretical result about the existence of perfect classifiers. Genetic Programming performs the evolution process of models, being responsible for exploring the search space of possible regressors and models. The results obtained with Econometric Genetic Programming – name of the algorithm in this dissertation – was compared with those from benchmarks in eight distinct cross-sectional datasets, showing competitive results in terms of accuracy in most cases. Both in the field of Genetic Programming and in that of econometrics, Econometric Genetic Programming has shown benefits such as help on introns identification, combat to bloat by statistical significance and generation of high level accuracy models, among others.

Keywords

Genetic Programming; Econometrics in Cross-Sectional Datasets; Regression and Classification.

Sumário

Lista de Figuras

Lista de tabelas

"É fazendo que se aprende a fazer
aquilo que se deve aprender a fazer"

Aristóteles

1
Introdução

1.1
Motivação e Objetivo

Quantificar a relação entre uma variável de resposta e um grupo de variáveis de controle é um dos problemas mais importantes da estatística (Denison et al., 2002). Esta tarefa geralmente se desdobra em duas: regressão e classificação. Ambas têm em comum a estimação numérica de uma função da variável de resposta e as variáveis de controle. A principal diferença entre as tarefas reside essencialmente no tipo de variável de resposta: se real, trata-se de uma tarefa de regressão; se categórica, trata-se de uma tarefa de classificação.

Considerando que alguns estudiosos apontam o ano de 1663 como o marco inicial da estatística (Willcox, 1938), com a publicação do trabalho *Natural and Political Observations upon the Bills of Mortality*, de John Graunt, pode-se dizer que as tarefas de regressão e classificação são tão antigas quanto a própria estatística.

Ao longo da história, no ramo da estatística, muitos foram os algoritmos que se propuseram a regredir e classificar uma variável de resposta. Ao buscar definir um algoritmo estado da arte para as tarefas, é mais coerente que se pense em termos do conjunto de dados que possui a variável de resposta a que se deseja estimar, do que em termos de um único algoritmo que apresente desempenho superior ao de todos os outros algoritmos em todos os conjuntos de dados – a existência desse algoritmo remeteria ao conceito de classificador perfeito (Davidson & MacKinnon, 2003). Não é possível definir um algoritmo que seja o estado da arte único para todos os conjuntos de dados. Em situações práticas, haverá grupos de algoritmos que tendem a apresentar melhor desempenho em determinados conjuntos de dados, em função das características do conjunto de dados e dos próprios algoritmos.

Quantificar a relação entre uma variável de resposta e um grupo de variáveis de controle não é uma tarefa exclusiva da estatística – ela reside em

diversas áreas do conhecimento, como, por exemplo, na ciência da computação. Nesse domínio, a computação evolucionária, sub-ramo da inteligência computacional, compreende um conjunto de metodologias computacionais e abordagens inspiradas em elementos da natureza, como a seleção natural, para resolver problemas diversos, tais como os de regressão e de classificação.

Na ciência da computação, assim como na estatística, há muitos algoritmos capazes de resolver tais tarefas – a Programação Genética (PG), técnica sistemática da computação evolucionária que automaticamente soluciona problemas sem a necessidade de se conhecerem informações do domínio do conjunto de dados, do problema ou formato da solução do problema, é um destes algoritmos.

Davidson et al. (1999) e Giustolisi & Savic (2006) estão entre os primeiros a utilizar ferramentas estatísticas e de computação evolucionária para resolver tarefas de regressão. Em linhas gerais, ambos combinam os processos estatísticos de estimação de parâmetros com o poder combinatório e de geração de modelos distintos da computação evolucionária – os algoritmos de regressão resultantes desta combinação apresentam maior capacidade preditiva do que a apresentada pelas mesmas técnicas quando utilizadas separadamente. Entretanto, as abordagens de Davidson et al. (1999) e Giustolisi & Savic (2006), embora aplicáveis a problemas práticos, carecem da apresentação de fundamentação teórica necessária ao pleno uso das ferramentas econométricas que propõem, além do fato de não serem aplicáveis a tarefas de classificação.

Esta dissertação tem como objetivo principal propor algoritmos de regressão e classificação de elevada acurácia, competitivos frente a algoritmos existentes que desempenhem as mesmas tarefas, através da utilização de ferramental estatístico e de computação evolucionária. Sua motivação surge da possibilidade de se utilizarem duas das características mais vantajosas de cada uma das vertentes: estimação de parâmetros – da estatística – e poder combinatório de geração de modelos distintos – da computação evolucionária.

A PG é a ferramenta mais indicada para gerar modelos distintos porque não somente cumpre com a geração de possíveis **regressores**, (o conceito de “regressor” será devidamente explicado posteriormente) permitindo aos modelos aproveitamento considerável do espaço de busca, como também realiza todo o

processo de evolução de modelos, utilizando operadores genéticos para propor novas regressões.

Há também, na construção desta dissertação, a motivação de prosseguir com o desenvolvimento de algoritmos que possam ser denominados classificadores perfeitos em uma ampla gama de conjuntos de dados – a noção de "classificador perfeito", embora aplicada à tarefa de classificação, se estende para a tarefa de regressão.

Os algoritmos propostos nesta dissertação possuem características peculiares e situações específicas nas quais podem ser aplicados, abordadas na seção 1.2.

1.2
Fundamentos

1.2.1
Tipos de Dados

As tarefas de regressão e classificação podem ser realizadas em distintos tipos de conjuntos de dados. Wooldridge (2008) afirma que dados de natureza econômica podem variar em sua estrutura – experimentalmente, observa-se que tal fato ocorre para dados de distintos campos do conhecimento, não somente da *econometria*.

A *econometria* consiste na aplicação da estatística matemática a conjuntos de dados econômicos, de forma a obter suporte empírico aos modelos construídos pela economia matemática e resultados numéricos (Tintner, 1968). Haavelmo (1944) acrescenta que tais resultados numéricos seriam alcançados pela utilização da inferência estatística como ferramenta. Embora se considere que o termo econometria tenha sido proposto por Ragnar Frisch e Jan Tinbergen – ganhadores do prêmio Nobel de economia em 1969 – como uma união entre estatística, matemática e teoria econômica, as ferramentas econométricas também podem ser utilizadas em conjuntos de dados de natureza qualquer, respeitando-se as categorias propostas. Portanto, nesta dissertação, o ferramental econométrico será utilizado independentemente da natureza (econômica, física, de engenharias etc.) do conjunto de dados.

Wooldridge (2006) fornece um panorama completo dos possíveis tipos de conjuntos de dados comumente encontrados em aplicações econométricas: de corte seccional (ou transversal), de séries de tempo, transversais agrupados ou de painel (ou longitudinais).

Dados de corte seccional ou transversal se caracterizam por terem sido coletados em um mesmo intervalo de tempo (diário, semanal, mensal, anual, etc.). Mesmo que isto não ocorra, considera-se que assim o seja e tal consideração é consistente, pois a diferença de tempo na coleta dos dados não interfere em sua característica. Por exemplo, se uma coleta de dados é feita com frequência anual e, para um determinado ano, parte dos dados tenha sido coletada na 1ª semana do ano e outra parte, na 2ª semana do ano, tal diferença temporal não é levada em consideração visto que o período de uma semana, tomando como referência o ano, é irrelevante para alterar qualquer característica dos dados.

Frequentemente, considera-se que dados seccionais são coletados como uma amostra aleatória da população de referência. Segundo Casella & Berger (2011), as variáveis aleatórias $y_1, y_2, \dots, y_t, \dots, y_n$ são chamadas de Amostra Aleatória (AA) de tamanho n da população $f(y)$ se $y_1, y_2, \dots, y_n$ forem variáveis aleatórias mutuamente independentes e a fdp (função densidade de probabilidade) ou fp (função de probabilidade) marginal de cada y_t, $t \in [1, n]$, for a mesma função $f(y)$. De modo alternativo, $y_1, y_2, \dots, y_n$ são chamadas variáveis aleatórias independentes e identicamente distribuídas (abrevia-se iid), com fdp ou fp igual à $f(y)$. Casella & Berger (2011), Spanos (1999) ou Hines et al. (2003) apresentam um tratamento adequado aos conceitos de variável aleatória, valor esperado, variância, fdp e fp. As siglas fdp, fp e iid podem ser apresentadas em caixa alta ou caixa baixa, dependendo da autoria do documento – nesta dissertação, serão apresentadas em caixa baixa, seguindo Casella & Berger (2011).

A partir da abordagem de Goldberger (1991), conclui-se que a hipótese de amostragem aleatória simplifica a análise de dados seccionais. Há algumas situações em que pode não ser adequado tratar um conjunto de dados transversal como AA. Independentemente de qual situação seja essa, frequentemente é a hipótese de independência a violada, embora a violação também possa ocorrer com a hipótese de distribuição idêntica entre $y_1, y_2, \dots, y_n$. A hipótese de amostragem aleatória mantém-se para todos os conjuntos de dados tratados nesta

dissertação, sendo não somente um efeito simplificador mas um direcionamento consistente com a natureza dos conjuntos de dados tratados.

Dados de séries de tempo consistem em observações de uma variável ou conjunto de variáveis ao longo do tempo. Duas diferenças fundamentais em relação aos dados transversais: o tempo é uma grandeza relevante, de tal modo que a ordem dos dados interfere na informação que se pode extrair de seu conjunto, e $y_1, y_2, \dots, y_n$ são geralmente dependentes. Além disso, a frequência de coleta de dados deve ser um fator de controle mais rígido neste tipo de conjunto de dados e pode haver a presença de tendências e/ou sazonalidade em séries temporais.

Dados (de corte) transversais agrupados possuem características dos dois tipos de conjuntos de dados anteriormente citados. Um conjunto de dados dessa natureza é construído da seguinte forma: coletam-se variáveis em um dado instante (preferencialmente em formato de AA), relativo a um grupo de controle e, em seguida, coletam-se as mesmas variáveis em um instante posterior, relativo a um novo grupo de controle. Nos dados transversais agrupados, o aspecto seccional está presente na coleta das mesmas variáveis; o aspecto temporal está presente na coleta de dados em dois momentos diferentes, mesmo que de grupos de controle distintos. Dados transversais agrupados são úteis quando o conjunto de dados em determinado instante de coleta é insuficiente em número e/ou deseja-se avaliar a eficácia de uma política de ações entre dois instantes.

Um conjunto de dados de painel (ou longitudinais) consiste em uma série de tempo para cada membro do corte transversal do conjunto – cada unidade de corte (indivíduo, empresa, elemento, etc.) possui a sua própria série temporal. A diferença evidente entre os dados longitudinais e os dados transversais agrupados é o fato de que, no primeiro, as mesmas unidades de corte transversal são acompanhadas ao longo do tempo. No segundo, as unidades não necessariamente serão as mesmas e, usualmente, não são.

Esta dissertação tem como foco o uso de regressão e classificação em conjuntos de dados seccionais, que formam a base para o estudo em dados transversais agrupado, de painel e de séries temporais, particularmente os modelos $AR\ (p)$ – facilmente derivados de modelos bem definidos para dados seccionais. Portanto, o intuito da dissertação é desenvolver satisfatoriamente modelos para

regressão e classificação em dados transversais para que, numa extensão deste trabalho, dados de outros tipos possam ser contemplados.

1.2.2
Linearidade e Não Linearidade

Hedrick & Girard (2010) afirma que um sistema linear deve satisfazer duas propriedades: superposição e homogeneidade. O princípio da superposição diz que, para duas entradas distintas, y_t e y_w, dentro do domínio da função $f(y)$, o resultado de $f(y_t + y_w)$ é igual à $f(y_t) + f(y_w)$. O princípio da homogeneidade afirma que, para qualquer número real k, $f(ky_t) = k\, f(y_t)$, para y_t pertencente ao domínio de $f(y)$.

Mardia et al. (1980) afirmam que o modelo linear geral é um modelo estatístico que pode ser escrito sob a forma $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$. A matriz $\boldsymbol{Y}$ é tal que $\boldsymbol{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_{i_0} \dots \boldsymbol{y}_m]$ é n x m, sendo n o número de observações de m distintas variáveis aleatórias correspondentes às colunas $\boldsymbol{y}_{i_0}, i_0 \in [1, m]$, com cada uma das m colunas $\boldsymbol{y}_{i_0}$ de $\boldsymbol{Y}$ sendo um vetor de dimensões n x 1. As variáveis aleatórias correspondentes às colunas $\boldsymbol{y}_{i_0}$ são nomeadas variáveis dependentes ou regressandos. Gujarati (2008) define variáveis dependentes como as variáveis que aparecem do lado esquerdo da igualdade $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$.

$\boldsymbol{X} \equiv [\boldsymbol{x}_1 \dots \boldsymbol{x}_k]$ é uma matriz de números nomeada matriz de regressores ou variáveis independentes e tem dimensões n x k, sendo k o número de regressores e cada coluna $\boldsymbol{x}_i, i \in [1, k]$, tem dimensão n x 1. Gujarati (2008) define variáveis independentes como as variáveis que aparecem do lado direito da igualdade $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$, à exceção de $\boldsymbol{\beta}$ e $\boldsymbol{U}$.

$\boldsymbol{\beta}$ é o vetor de parâmetros (ou coeficientes) que ajusta $\boldsymbol{X}$ a $\boldsymbol{Y}$. O vetor $\boldsymbol{\beta}$ tem dimensões k x m e, usualmente, seu estimador $\widehat{\boldsymbol{\beta}}$ é dado por um método de otimização. É importante atentar que $\boldsymbol{\beta}$ não necessariamente faz um ajuste perfeito de $\boldsymbol{X}$ a $\boldsymbol{Y}$. A matriz $\boldsymbol{U}$, nomeada matriz de erros e de dimensões n x m, compõe o modelo linear geral representando numericamente o quanto de $\boldsymbol{X}$ que não se ajusta a $\boldsymbol{Y}$ por meio de $\boldsymbol{\beta}$.

Se $\boldsymbol{Y}$ for um vetor coluna composto de somente um $\boldsymbol{y}_{i_0}$, de tal forma que $\boldsymbol{Y}$ seja nomeado $\boldsymbol{y}$, $\boldsymbol{\beta}$ for uma matriz k x 1 e $\boldsymbol{U}$ for n x 1, de tal forma que $\boldsymbol{U}$ seja

nomeado $\boldsymbol{u}$, então o modelo linear geral é nomeado modelo de regressão linear múltipla e assume a forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, que atende às propriedades de superposição e homogeneidade e resulta em um sistema de equações lineares. Sob circunstâncias que serão tratadas no capítulo seguinte, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ tem solução única.

Segundo Davidson & MacKinnon (1993), é comum que aplicações em econometria para resolução de problemas práticos elaborem modelos que busquem o estimador de $\boldsymbol{y}$ em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, e não o estimador de $\boldsymbol{Y}$ em $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}$. Esta dissertação realiza a mesma prática: deseja-se estimar a relação entre uma única variável dependente e um conjunto de variáveis independentes.

Qualquer sistema ou função que não obedeça às propriedades de superposição e homogeneidade é não linear – não há conjunto de características que sistemas ou funções devam obedecer para serem não lineares: basta que não sejam lineares (Hedrick & Girard, 2010).

Wooldridge (2001) afirma que um problema/sistema não linear é semelhante a um problema/sistema no qual as variáveis não têm forma fechada – ou seja, não há solução única para o sistema. Para estes problemas, algoritmos iterativos de distintas naturezas são necessários.

1.2.3
Princípio da Parcimônia

O princípio da parcimônia, ou navalha de Occam, é frequentemente considerado um dos princípios fundamentais da ciência moderna (Domingos, 1999). A navalha de Occam postula que as entidades não deveriam se multiplicar além da necessidade. Segundo Myung e Pitt (1997), sob o domínio de seleção de modelos, o objetivo ao se aplicar o princípio é escolher o modelo mais simples que seja capaz de descrever suficientemente bem o conjunto de dados.

Em outras palavras, o princípio afirma que a explicação para qualquer fenômeno deve considerar apenas as premissas e entidades estritamente necessárias à explicação do mesmo e eliminar todas as que não causariam qualquer diferença nas predições do fenômeno.

Em tarefas de regressão e classificação, o princípio se aplica da seguinte forma: deseja-se regredir/classificar uma variável de resposta em função do menor

número possível de variáveis independentes, preferencialmente a partir de um modelo que tenha a estrutura mais simples possível.

1.2.4
Modelos Caixa-branca, Caixa-preta e Caixa-cinza

Ljung (1999) e Giustolisi (2004) afirmam que é usual que cores sejam utilizadas para categorizar modelos matemáticos em função do nível de informação a priori requisitada pelos modelos.

Giustolisi & Savic (2006) detalham estas categorias. Modelos de caixa-branca são definidos como sistemas onde toda a informação necessária ao pleno entendimento do experimento de interesse está disponível. Isto é, o modelo é baseado em princípios primários (por exemplo, modelos associados a leis físicas), variáveis e parâmetros já conhecidos. Nos modelos de caixa-branca, pelo fato de variáveis e parâmetros terem um significado/interpretação, estes também explicam as relações existentes entre as distintas partes do sistema.

Modelos caixa-preta são modelos nos quais não há informação disponível a priori, sendo essa uma de suas maiores potencialidades, pois fornece ao seu usuário informação de um sistema do qual o próprio usuário o desconhece ou conhece insuficientemente. Tais modelos são orientados ao conjunto de dados aos quais são submetidos, buscando descobrir a forma funcional (ou seja, a relação entre variável dependente e as variáveis independentes) e os respectivos parâmetros provenientes da forma funcional, que usualmente precisam ser estimados.

Modelos caixa-cinza possuem sua estrutura matemática derivada de princípios primários (de fenômenos físicos ou de simplificações de equações diferencias descrevendo o fenômeno, por exemplo). Geralmente há necessidade de estimação de parâmetros por meio do conjunto de dados disponível, embora o intervalo numérico esperado dos parâmetros estimados seja conhecido.

1.3
Descrição do Trabalho

Os fundamentos das seções anteriores permitem caracterizar de maneira mais completa o objetivo desta dissertação. Como dito anteriormente, há o objetivo principal de propor modelos de regressão e classificação de elevada acurácia, competitivos frente a algoritmos que desempenhem as mesmas tarefas. Os modelos propostos serão do tipo caixa-preta, aplicados exclusivamente a conjuntos de dados seccionais, mantendo a hipótese de amostragem aleatória para todos os conjuntos de dados tratados.

Os modelos de regressão são lineares. Os modelos de classificação, embora não lineares, têm uma porção de sua estrutura que é linear, permitindo leitura simples das variáveis independentes que compõem o modelo. Ambos os modelos permitem interpretabilidade dos parâmetros estimados, embora este não seja o objetivo principal da dissertação.

Os modelos de regressão e classificação obedecem ao princípio da parcimônia, através da seleção de variáveis estatisticamente significantes (vide seção 2.3 e capítulo 4).

Estruturou-se esta dissertação nas etapas listadas a seguir, como uma possível forma de se obter algoritmos dentro das características citadas.

(1) Definição dos modelos de regressão e classificação: os modelos de regressão têm a forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ por esta apresentar solução única sob algumas condições. Os modelos **logit** serão utilizados para classificação binária, devido essencialmente ao fato de estimarem a probabilidade de determinado observação pertencer a uma dada classe e pela sua forte difusão como algoritmo de classificação em econometria.

(2) Definição dos algoritmos que solucionam a estimação de $\widehat{\boldsymbol{\beta}}$ nos modelos de regressão e classificação: o algoritmo que soluciona a estimação de $\widehat{\boldsymbol{\beta}}$, nos modelos de regressão, é a decomposição QR e, nos modelos de classificação, é uma variante do Método de Newton. Ao passo que a decomposição QR lida de maneira satisfatória com multicolinearidade e fornece soluções numéricas precisas, o Método de Newton atinge o ótimo global de problemas de otimização sob determinadas circunstâncias.

(3) Estudo das condições necessárias para realização de Testes de Hipóteses: são plenamente satisfeitas quando n – o número de observações do conjunto de dados – é suficientemente grande ou $n \to \infty$, sob condições adicionais avaliadas posteriormente, permitindo que a estatística de teste em questão tenha distribuição assintótica conhecida.

(4) Apresentação da PG como a ferramenta que realiza o processo de evolução e geração de modelos. A PG se utilizará de algumas ferramentas para que possa explorar o espaço de busca de maneira ampla. São elas: representação por árvores, por possuir grande potencial de geração de distintos regressores para os modelos; utilização dos operadores de mutação e cruzamento entre indivíduos. A PG fará uso das funções de avaliação Raiz do Erro Quadrático Médio (para tarefas de regressão) e Percentual de Classificações Incorretas (para tarefas de classificação).

(5) Introdução e motivação, sob aspectos teóricos, do algoritmo gerador de modelos de regressão e classificação.

(6) Foram propostos oito estudos de caso – cinco para regressão e três para classificação – para avaliar o desempenho dos modelos.

Cita-se como trabalhos diretamente relacionados a esta dissertação Davidson et al. (1999) e Giustolisi & Savic (2006).

Davidson et al. (1999) propõem modelos lineares, com estimação de $\widehat{\boldsymbol{\beta}}$ por MQO (conforme seção 2.2.1) e TH (vide seção 2.3) em $\boldsymbol{\beta}$ via *Backward Elimination* (veja Draper & Smith (1998)) para tarefas de regressão (e não de classificação).

Giustolisi & Savic (2006) propõem o algoritmo de EPR (*Evolutionary Polynomial Regression*) para solucionar tarefas de regressão (e não de classificação). A EPR utiliza um algoritmo genético para explorar o espaço de busca de modelos e MQO para estimar $\widehat{\boldsymbol{\beta}}$; minimiza uma função objetivo que penaliza por acréscimo de regressores e realiza TH sobre $\boldsymbol{\beta}$.

1.4
Organização da Dissertação

Os próximos capítulos estão organizados como segue. O capítulo 2 apresenta os fundamentos econométricos necessários à construção dos modelos de regressão e classificação. O capítulo 3 apresenta a Programação Genética, ferramenta da computação evolucionária que realiza o processo de geração de modelos. O capítulo 4 apresenta em detalhes o mecanismo gerador de modelos, fundamentando sua razão teórica e os elementos da computação evolucionária necessários à sua construção. O capítulo 5 apresenta os experimentos e seus resultados. No capítulo 6 realiza-se a conclusão da dissertação, junto a propostas de trabalhos futuros, avaliação de benefícios e limitações do algoritmo proposto.

2
Econometria

Esta dissertação gera modelos lineares para conjuntos de dados associados à tarefa de regressão e modelos não lineares para conjuntos de dados associados à tarefa de classificação. A apresentação dos fundamentos econométricos que sustentam cada grupo de modelos será feita separadamente. Tais fundamentos estão ligados essencialmente à forma do modelo e processo de estimação. A terceira parte deste capítulo aborda testes de hipóteses, ferramenta indispensável à natureza parcimoniosa dos modelos.

O capítulo dois se utilizará de uma quantidade significativa de equações e fórmulas – resultados teóricos importantes dentro do domínio da econometria. Para efeito de simplificação e fluência do texto, quando em referência a uma equação ou fórmula, não será utilizado o termo "eq." e "form.", justamente pela chamada a equações e fórmulas ser feita com frequência.

2.1
Modelos de Regressão Linear

Modelos de Regressão linear são modelos da forma:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{2.1}$$

onde $\boldsymbol{y}_{n\,\mathrm{x}\,1}$ é o vetor de n observações da variável dependente y, tal que $\boldsymbol{y} = [y_1\ y_2\ \cdots\ y_n]^{\mathrm{T}}$; $\boldsymbol{X}_{n\,\mathrm{x}\,k}$ é a matriz de n observações das k variáveis independentes $X = \{x_1, x_2, \ldots, x_k\}$, de tal forma que $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \ldots \boldsymbol{x}_i \ldots \boldsymbol{x}_k]$, em que cada coluna $\boldsymbol{x}_i, i \in [1, k]$, é um vetor n x 1; $\boldsymbol{\beta}_{k\,\mathrm{x}\,1}$ faz o ajuste de $\boldsymbol{X}$ a $\boldsymbol{y}$ e, por reconhecer-se que este ajuste pode não ser perfeito, inclui-se em (2.1) o termo de erro $\boldsymbol{u}_{n\,\mathrm{x}\,1} = [u_1\ u_2\ \cdots\ u_n]^{\mathrm{T}}$.

É comum que seja adicionada a $\boldsymbol{X}$ uma coluna unitária e, à $\boldsymbol{\beta}$, um termo extra, denominado intercepto. Desta forma, $\boldsymbol{X}_{n\,\mathrm{x}\,(k+1)}$ e $\boldsymbol{\beta}_{(k+1)\,\mathrm{x}\,1}$. A seção

seguinte proporá um arcabouço teórico que não considera explicitamente tais adições a $\boldsymbol{X}$ e $\boldsymbol{\beta}$, o que não torna o arcabouço teórico inválido.

Usualmente, n é uma parcela das observações que se pode obter em uma situação real de amostragem aleatória, fazendo com que nenhum dos elementos em (2.1) seja calculável, a não ser que n seja o número de todas as observações que se possa obter de y e X – o que frequentemente não ocorre.

(2.1) é nomeado modelo populacional. O conjunto de n observações de y e X constitui uma amostra numérica destas variáveis. Tal amostra é suficiente, sob certas condições, para que se possa dizer algo a respeito da relação populacional entre y e X.

$\boldsymbol{\beta}$ é o elemento capaz de informar sobre a relação existente entre y e X. Todavia, $\boldsymbol{\beta}$ também é desconhecido, por ser um parâmetro populacional. Deve-se, portanto, utilizar a amostra de n observações para que se possa produzir alguma informação relacionada à $\boldsymbol{\beta}$, através de seu estimador, $\widehat{\boldsymbol{\beta}}$.

2.1.1
Estimação por Mínimos Quadrados Ordinários

Esta seção é baseada em Davidson & MacKinnon (1993).

O método de estimação mais utilizado em econometria é o método de Mínimos Quadrados (MQ). Em função da linearidade, há dois possíveis métodos de MQ: Mínimos Quadrados Ordinários (MQO) – a equação de regressão que será estimada é linear nos parâmetros – e Mínimos Quadrados Não Lineares (MQNL) – a equação é não linear em pelo menos um dos parâmetros. A estimação de MQO pode ser calculada de diversas maneiras, todas diretas (não iterativas), enquanto a estimação por MQNL requer procedimentos iterativos. Esta seção trata exclusivamente do método de MQO.

Há uma importante distinção a se fazer entre as propriedades numéricas e estatísticas da estimação por MQO. Propriedades numéricas são aquelas derivadas do uso do MQO, independentemente de como o conjunto de dados foi gerado – todas as propriedades numéricas do MQO podem ser interpretadas em termos de geometria euclidiana. Propriedades estatísticas, para que se mantenham, têm como base certas hipóteses sobre a forma como o conjunto de dados foi gerado. Tais

hipóteses jamais podem ser verificadas de tal forma que retornem uma resposta inquestionável com relação à sua veracidade – entretanto, em alguns casos, elas podem ser testadas via testes de hipóteses. Esta seção apresenta o MQO como ferramenta matemática e computacional, sem introduzir formalmente qualquer tipo de propriedade estatística que a suporte.

Os componentes essenciais de uma regressão linear são o regressando $\boldsymbol{y}$ e a matriz de regressores $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \dots \boldsymbol{x}_i \dots \boldsymbol{x}_k]$. Quando se fizer referência ao índice i, presume-se que $i \in [1, k]$ e tal informação será omitida, assim como se pode referir a $\boldsymbol{X}$ como $[\boldsymbol{x}_1 \dots \boldsymbol{x}_k]$. A matriz $\boldsymbol{y}$ e cada um dos $\boldsymbol{x}_i$ podem ser pensados como pontos em um espaço euclidiano n-dimensional, E^n. Supondo que sejam linearmente independentes, os k regressores formam um subespaço k-dimensional de E^n, nomeado $\delta(\boldsymbol{X}) = \delta(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$, que possui dimensão sempre igual ao posto de $\boldsymbol{X}$, nomeado $\rho(\boldsymbol{X})$. Por definição, $\delta(\boldsymbol{X})$ consiste de todos os pontos $\boldsymbol{z}$ em E^n tal que $\boldsymbol{z} = \boldsymbol{X\gamma}$, para algum $\boldsymbol{\gamma}_{k \text{ x } 1}$. Considera-se que $n \gg k$, o que é usual em casos práticos de econometria, o que torna possível que $\boldsymbol{X}$ tenha posto cheio (os $\boldsymbol{x_i}$ são todos linearmente independentes), igual à k.

Se um ponto $\boldsymbol{z}$, um vetor n x 1, pertence a $\delta(\boldsymbol{X})$, pode-se escrever $\boldsymbol{z}$ como uma combinação linear das colunas de $\boldsymbol{X}$.

$$\boldsymbol{z} = \boldsymbol{X\gamma} = \sum_{i=1}^{k} \gamma_i \boldsymbol{x}_i \tag{2.2}$$

onde γ_i são escalares que compõem o vetor $\boldsymbol{\gamma}$. Portanto, qualquer vetor de k coeficientes, como $\boldsymbol{\gamma}$ (que é k x 1), identifica qualquer ponto em $\delta(\boldsymbol{X})$. Se os $\boldsymbol{x_i}$ são linearmente independentes – ou seja, se não se pode escrever nenhum deles como combinação linear dos demais – (2.2) tem solução única. Deste parágrafo em diante, considera-se que as colunas de $\boldsymbol{X}$ são linearmente independentes.

Pode-se realizar qualquer transformação linear em $\boldsymbol{X}$, contanto que seja preservado o posto de $\boldsymbol{X}$, de tal forma que o subespaço gerado pela matriz $\boldsymbol{X}$ transformada é o mesmo que $\delta(\boldsymbol{X})$. Utilizando (2.2) e supondo $\boldsymbol{X}^* = \boldsymbol{XA}$, onde $\boldsymbol{A}$ representa a matriz da transformação linear:

$$\boldsymbol{z} = \boldsymbol{X}^{*}\boldsymbol{A}^{-1}\boldsymbol{\gamma} \equiv \boldsymbol{X}^{*}\boldsymbol{\gamma}^{*} \tag{2.3}$$

Portanto, assim como qualquer $\boldsymbol{z}$ pode ser escrito como combinação linear das colunas de $\boldsymbol{X}$, também é possível que qualquer $\boldsymbol{z}$ possa ser escrito como combinação linear de quaisquer transformações lineares das colunas de $\boldsymbol{X}$. Consequentemente, se $\delta(\boldsymbol{X})$ é gerado pelas colunas de $\boldsymbol{X}$, $\delta(\boldsymbol{X})$ também é gerado por $\boldsymbol{X}^{*} = \boldsymbol{X}\boldsymbol{A}$.

O complemento ortogonal de $\delta(\boldsymbol{X})$ em E^{n}, denotado por $\delta^{\perp}(\boldsymbol{X})$, é o conjunto de todos os pontos $\boldsymbol{w}$ em E^{n} tal que, para todo $\boldsymbol{z}$ em $\delta(\boldsymbol{X})$, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{0}$. Logo, todo ponto em $\delta^{\perp}(\boldsymbol{X})$ é ortogonal a todo ponto em $\delta(\boldsymbol{X})$ – dois pontos são ditos ortogonais se o produto interno entre ambos é zero (Poole, 2010). Como a dimensão de $\delta(\boldsymbol{X})$ é k, a dimensão de $\delta^{\perp}(\boldsymbol{X})$ é $n - k$.

A Figura 2.1 ilustra os conceitos acima, para $n = 2$ e $k = 1$. A matriz $\boldsymbol{X}$ tem somente uma coluna nesse caso, sendo representada por um vetor. Consequentemente, $\delta(\boldsymbol{X})$ é unidimensional e, como $n = 2$, $\delta^{\perp}(\boldsymbol{X})$ também é unidimensional.

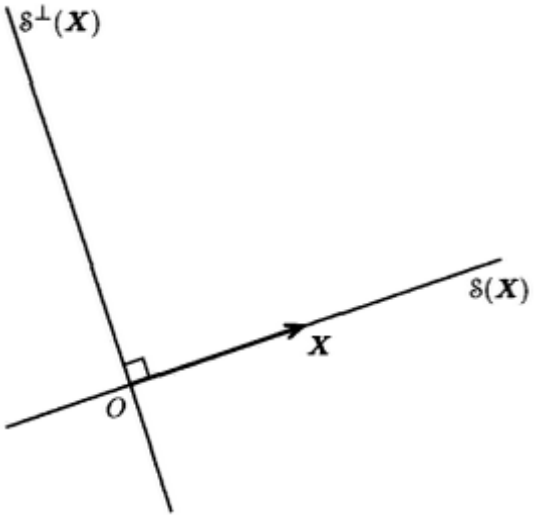

Figura 2.1 – Os espaços $\boldsymbol{\delta}(\boldsymbol{X})$ e $\boldsymbol{\delta}^{\perp}(\boldsymbol{X})$

Fonte: Davidson & MacKinnon (1993)

Como já pontuado, qualquer ponto em $\delta(\boldsymbol{X})$ pode ser representado por um vetor na forma $\boldsymbol{X\beta}$, para algum $\boldsymbol{\beta}_{k\,\mathrm{x}\,1}$. Se o objetivo for determinar o ponto em $\delta(\boldsymbol{X})$ que é o mais próximo possível de um vetor $\boldsymbol{y}$, o problema passa a ser minimizar a distância entre $\boldsymbol{y}$ e $\boldsymbol{X\beta}$, com respeito à $\boldsymbol{\beta}$. Minimizar essa distância é equivalente a minimizar o quadrado da distância, já que a função polinomial de grau dois é monotônica crescente (Ashlagi et al., 2010) para valores maiores ou iguais a zero e não há distâncias negativas. (2.4) expressa o problema matematicamente. O estimador de MQO, $\widehat{\boldsymbol{\beta}}$, é o valor que soluciona (2.4).

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X\beta}\|^2 = \min_{\boldsymbol{\beta}} \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X\beta}) \qquad (2.4)$$

onde y_t e $\boldsymbol{X}_t$ representam, respectivamente, o t-ésimo elemento de $\boldsymbol{y}$ e a t-ésima linha de $\boldsymbol{X}$. Ou seja, t representa uma evidência (pessoa, cidade, item ou semelhante) da AA coletada para a formação do conjunto de dados, de tal forma que $t \in [1, n]$. Quando se fizer referência ao índice t, presume-se que $t \in [1, n]$ e tal informação será omitida.

O vetor de resíduos $\widehat{\boldsymbol{u}}$ é dado por $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Embora $\boldsymbol{u}$ não seja mensurável, amostralmente é possível obter uma grandeza relativa ao erro da estimação, que é o próprio resíduo $\widehat{\boldsymbol{u}}$. Como o resíduo, observação a observação, é dado por $\hat{u}_t = y_t - \boldsymbol{X}_t\widehat{\boldsymbol{\beta}}$, $\sum_{t=1}^{n}(y_t - \boldsymbol{X}_t\widehat{\boldsymbol{\beta}})^2$ é nomeado Somatório dos Quadrados dos Resíduos (SQR).

A Figura 2.2 ilustra a geometria do método de MQO. O regressando é o vetor $\boldsymbol{y}$. O vetor $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ é ponto mais próximo de $\boldsymbol{y}$ em $\delta(\boldsymbol{X})$ – nesse caso, com $n = 2$ e $k = 1$, $\widehat{\boldsymbol{\beta}}$ é um escalar. O segmento de reta, que liga $\boldsymbol{y}$ à $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ e faz ângulo reto com $\delta(\boldsymbol{X})$ em $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$, é simplesmente o vetor $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ transladado da origem até o ponto $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

O ângulo reto formado entre $\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ e $\delta(\boldsymbol{X})$ é essencial na estimação por MQO. Em qualquer outro ponto de $\delta(\boldsymbol{X})$, como $\boldsymbol{X\beta}'$ (Figura 2.2), $\boldsymbol{y}-\boldsymbol{X\beta}'$ não forma ângulo reto com $\delta(\boldsymbol{X})$. Por consequência, $\|\boldsymbol{y}-\boldsymbol{X\beta}'\|$ deve ser necessariamente maior que $\left\|\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|$.

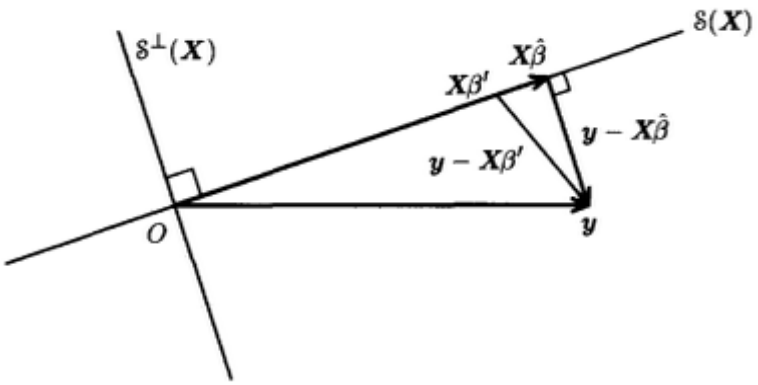

Figura 2.2 – A projeção de $\boldsymbol{y}$ em $\boldsymbol{\delta}(\boldsymbol{X})$

Fonte: Davidson & MacKinnon (1993)

O vetor de derivadas de (2.4) em relação aos elementos de $\boldsymbol{\beta}$ é:

$$-2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{2.5}$$

Esta equação (2.5) deve se igualar a zero em um ponto de mínimo. Como $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ tem posto cheio (pois foi considerado que as colunas de $\boldsymbol{X}$ são linearmente independentes) e qualquer matriz na forma $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ é necessariamente não negativa definida, o SQR é uma função estritamente convexa de $\boldsymbol{\beta}$ e, consequentemente, tem um ótimo único $\widehat{\boldsymbol{\beta}}$, determinado pelas equações em (2.6). Borwein & Lewis (2005) abordam os conceitos de convexidade estrita e formas de matrizes.

$$\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{0} \tag{2.6}$$

As equações em (2.6), nomeadas equações normais, expressam que $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ deve ser ortogonal a todo $\boldsymbol{x}_i$ e, por isso, a qualquer vetor que esteja contido em $\delta(\boldsymbol{X})$ – tal assertiva, do ponto de vista geométrico, é equivalente a dizer que $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ e $\delta(\boldsymbol{X})$ devem formar um ângulo reto entre si. Portanto, (2.6) expressa algebricamente o que a Figura 2.2 expressa geometricamente.

Como a matriz $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ tem posto cheio, sempre é possível invertê-la e resolver (2.6) para $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{2.7}$$

$\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ é único mesmo que $\boldsymbol{X}$ não tenha posto cheio, visto que, geometricamente, $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ é simplesmente o ponto em $\delta(\boldsymbol{X})$ mais próximo de $\boldsymbol{y}$. O mesmo não pode ser dito para $\widehat{\boldsymbol{\beta}}$: se $\boldsymbol{X}$ não tem posto cheio, o sistema (2.7) não tem solução única.

2.1.2
Decomposição QR

O objetivo desta seção é propor um algoritmo que calcule a estimativa de MQO e é baseada em Davidson & MacKinnon (1993).

O estimador de MQO, $\widehat{\boldsymbol{\beta}}$, é dado pela expressão (2.7). A estimativa de MQO é um vetor de números reais, obtido após todos os procedimentos de cálculo que envolvem $\widehat{\boldsymbol{\beta}}$ terem sido realizados.

Se $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ e $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ forem calculados, utilizando-se de uma rotina genérica de inversão de matrizes para inverter $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ e, em seguida, multiplicando-se o resultado por $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$, a estimativa de MQO é encontrada. Esta maneira, embora trivial, é numericamente instável. Tipicamente, um algoritmo de natureza instável não garante que os erros de aproximação diminuam conforme haja flutuações no conjunto de dados de entrada (Burden & Faires, 2011). A metodologia descrita acima pode funcionar satisfatoriamente se a precisão adequada na representação de dados for utilizada, as colunas de $\boldsymbol{X}$ forem similares em magnitude (é incomum em situações práticas, a não ser que algum método de normalização seja utilizado) e se $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ não estiver muito próximo de ser singular (não inversível).

Chambers (1977) e Maindonald (1984) recomendam a decomposição QR (ou fatoração QR) para o cálculo da estimativa – Chambers (1977) não recomenda outro método que não seja QR e Maindonald (1984) acrescenta que o método é recomendado em experimentos onde a acurácia é de extrema importância. Davidson & MacKinnon (1993) também a utilizam, argumentando que os

resultados alcançados através dela são mais precisos e que a metodologia é mais interessante do ponto de vista teórico.

A decomposição QR determina uma base ortonormal para $\delta(\boldsymbol{X})$ (o subespaço gerado pelas colunas de $\boldsymbol{X}$), nomeada $\boldsymbol{Q}_{n \text{ x } k}$, com as propriedades: $\delta(\boldsymbol{X}) = \delta(\boldsymbol{Q})$ e $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \mathbf{I}$. Por definição, uma base de E^n é dita ortonormal se esta mesma base é ortogonal – os vetores da base são ortogonais dois a dois, entre si – e todos os seus componentes são vetores unitários. Para qualquer $\boldsymbol{X}$ de posto cheio, é possível realizar uma decomposição QR determinando-se $\boldsymbol{Q}$ e $\boldsymbol{R}_{k \text{ x } k}$, sendo $\boldsymbol{R}$ uma matriz triangular superior, tal que:

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} \quad \text{e} \quad \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \mathbf{I} \tag{2.8}$$

A equação $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \mathbf{I}$ implica em ortonormalidade das colunas de $\boldsymbol{Q}$. A matriz $\boldsymbol{R}$ ser triangular implica no fato das colunas de $\boldsymbol{Q}$ serem geradas recursivamente: a 1ª coluna de $\boldsymbol{Q}$ é a 1ª coluna de $\boldsymbol{X}$, redimensionada para que tenha tamanho unitário; a 2ª coluna de $\boldsymbol{Q}$ é uma transformação linear das primeiras duas colunas de $\boldsymbol{X}$, que é ortogonal à 1ª coluna de $\boldsymbol{Q}$ e também possui tamanho unitário; e assim por diante até que $\boldsymbol{Q}$ tenha dimensões n x k. Quando o posto de $\boldsymbol{X}$ não é cheio e há m colunas de $\boldsymbol{X}$ que sejam linearmente dependentes das $k - m$ colunas restantes, o algoritmo é modificado de tal forma que $\boldsymbol{Q}$ tenha $k - m$ colunas e $\boldsymbol{R}$ tenha dimensões $(k - m)$ x k. Desta forma, a estimativa de $\widehat{\boldsymbol{\beta}}$ torna-se solução única do algoritmo quando se arbitra que os coeficientes dos m regressores linearmente dependentes sejam iguais à zero.

Tendo as matrizes $\boldsymbol{Q}$ e $\boldsymbol{R}$ calculadas pela fatoração QR, a função de regressão $\boldsymbol{X}\boldsymbol{\beta}$ pode ser escrita como $\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{\gamma}$. Estima-se, inicialmente, $\boldsymbol{\gamma}$ através de $\widehat{\boldsymbol{\gamma}} = (\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q})^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}$, que tem estrutura semelhante à (2.7), utilizando-se $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \mathbf{I}$, $\widehat{\boldsymbol{\gamma}} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y}$. Geometricamente, as Figuras 2.1 e 2.2 seriam as mesmas se fosse feita a substituição de $\boldsymbol{X}$ por $\boldsymbol{Q}$ como matriz de regressores, pois $\delta(\boldsymbol{X}) = \delta(\boldsymbol{Q})$.

Para o cálculo da estimativa de $\widehat{\boldsymbol{\beta}}$, $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ e $\widehat{\boldsymbol{u}}$ utilizam-se (2.9), (2.10) e (2.11), respectivamente.

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R}^{-1}\widehat{\boldsymbol{\gamma}} \tag{2.9}$$

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = (\boldsymbol{R}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{R})^{-1} = (\boldsymbol{R}^{\mathrm{T}}\mathbf{I}\boldsymbol{R})^{-1} = (\boldsymbol{R}^{\mathrm{T}}\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})^{\mathrm{T}} \tag{2.10}$$

$$\widehat{\boldsymbol{u}} = \boldsymbol{y} - \boldsymbol{Q}\widehat{\boldsymbol{\gamma}} = \boldsymbol{y} - \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y} \tag{2.11}$$

A decomposição QR realiza internamente toda a rotina de cálculos para $\widehat{\boldsymbol{\beta}}$ e outras grandezas associadas ao erro, que serão apresentadas posteriormente.

Como $\boldsymbol{R}$ é triangular, $\boldsymbol{R}^{-1}$ é facilmente calculada, sendo uma operação pouco custosa para o algoritmo. Não há nem mesmo a necessidade de verificar singularidade para $\boldsymbol{R}$, pois esta não terá posto cheio se $\boldsymbol{X}$ também não o tiver – utilizam-se aqui as definições de $\boldsymbol{X}$, $\boldsymbol{Q}$ e $\boldsymbol{R}$, além das relações que têm entre si.

A parte mais custosa da fatoração é formar $\boldsymbol{Q}$ e $\boldsymbol{R}$ de $\boldsymbol{X}$ – a complexidade desta é operação é proporcional a nk^2. A formação da matriz de somas quadráticas e produtos cruzados, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, que é o 1º passo para métodos baseados em soluções de equações normais, também tem complexidade proporcional a nk^2, embora o fator de proporcionalidade seja menor.

Há alguns algoritmos para cálculo da decomposição QR: algoritmo de Gram-Schmidt (Björck, 1967, e Ling et al., 1986), processo de transformações de Householder (Businger & Golub, 1965, e Hanson & Lawson, 1969), rotações de Givens (Ling, 1991), dentro outros. Os experimentos desta dissertação são realizados no *software Matlab* R2014a. A rotina para cálculo da decomposição QR não é disponibilizada pela *Mathworks*, empresa que desenvolve o *Matlab*.

2.2
Modelos de Regressão Não Linear: Classificação Binária

De acordo com Davidson & MacKinnon (1993), a classificação binária é o caso mais comum em situações práticas. Neste tipo de modelo, o valor da variável dependente y_t só pode assumir dois valores, 1 ou 0 – tais valores indicam se um evento de interesse ocorreu ou não. Supondo $y_t = 1$ o indicativo de que o evento ocorreu para a observação t e $y_t = 0$ o indicativo para o caso contrário, pode-se considerar uma probabilidade (condicional) de o evento ter ocorrido, nomeada P_t.

O modelo de classificação binária tem como objetivo modelar P_t condicional a $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \dots \boldsymbol{x}_k]$ ou a $X = \{x_1, x_2, \dots, x_k\}$: tanto $\boldsymbol{X}$ quanto X são adequados nesse caso específico. Será $\boldsymbol{X}$ a notação utilizada. Matematicamente, expressa-se P_t como o valor esperado de y_t condicional a $\boldsymbol{X}$:

$$P_t \equiv \Pr(y_t = 1|\boldsymbol{X}) = E(y_t|\boldsymbol{X}) \tag{2.12}$$

A partir de (2.12), conclui-se que o modelo de regressão linear não é adequado para tarefas de classificação binária. Para chegar a tal conclusão, propõe-se que $\boldsymbol{X}_t$ seja um vetor linha (dimensões $1 \times k$) de variáveis que pertençam a X, incluindo uma constante para efeito de cálculo. Dessa forma, um modelo de regressão linear especificaria $E(y_t|\boldsymbol{X})$ como $\boldsymbol{X}_t\boldsymbol{\beta}$. Entretanto, $E(y_t|\boldsymbol{X})$ é uma probabilidade, por definição, e probabilidades necessariamente devem assumir valores entre 0 e 1, inclusive, de acordo com os axiomas de Kolmogorov (Kolmogorov, 1960). Como a grandeza $\boldsymbol{X}_t\boldsymbol{\beta}$ não é restrita a qualquer conjunto fechado de valores reais, $\boldsymbol{X}_t\boldsymbol{\beta}$ não pode ser interpretada como probabilidade – por conseguinte, inviabilizando o modelo de regressão linear para tarefas de classificação binária.

Há vários modelos disponíveis para modelar (2.12) e, por consequência, ser utilizados para classificação – os mais amplamente divulgados são os modelos **probit** e **logit**, de acordo com Fernandes (2009). Ambos consistem em uma função de transformação, $F(x)$, aplicada a uma função índice, $h(\boldsymbol{X}_t, \boldsymbol{\beta})$, que depende exclusivamente de $\boldsymbol{X}_t$ e $\boldsymbol{\beta}$. A função índice $h(\boldsymbol{X}_t, \boldsymbol{\beta})$ tem as propriedades de uma função de regressão, seja ela linear ou não linear. A especificação geral de

um modelo para classificação, que modela probabilidades condicionais é dada por:

$$E(y_t|\boldsymbol{X}) = F(h(\boldsymbol{X}_t, \boldsymbol{\beta})) \tag{2.13}$$

Uma maneira mais restritiva, entretanto mais comum, de modelar P_t é:

$$E(y_t|\boldsymbol{X}) = F(\boldsymbol{X}_t\boldsymbol{\beta}) \tag{2.14}$$

Por construção, $F(x)$ tem as propriedades (2.15) e (2.16), fazendo com que a transformação linear seja uma função monotonicamente crescente que mapeie da reta real ao intervalo 0-1.

$$F(-\infty) = 0 \quad \text{e} \quad F(\infty) = 1 \tag{2.15}$$

$$\frac{\partial F(x)}{\partial x} > 0 \tag{2.16}$$

Em (2.14), a função índice $\boldsymbol{X}_t\boldsymbol{\beta}$ é linear e $E(y_t|\boldsymbol{X})$ é uma transformação não linear de $\boldsymbol{X}_t\boldsymbol{\beta}$. Embora $\boldsymbol{X}_t\boldsymbol{\beta}$ possa assumir, em princípio, qualquer valor real, $F(\boldsymbol{X}_t\boldsymbol{\beta})$ somente pode assumir valores entre 0 e 1, devido a (2.15).

Há várias funções cumulativas de distribuições que possuem as propriedades (2.15) e (2.16), fazendo com que se possa modelar P_t de distintas maneiras. Em situações práticas, (2.14) é quase sempre preferível à (2.13) – a função de regressão linear $\boldsymbol{X}_t\boldsymbol{\beta}$ é quase sempre preferível a uma função de regressão não linear genérica $h(\boldsymbol{X}_t, \boldsymbol{\beta})$ – sendo os modelos probit e logit os mais utilizados. Os modelos diferem com relação à especificação de $F(\cdot)$.

Para o modelo probit, $F(x)$ é a função cumulativa da distribuição normal, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}X^2\right) dX$, de tal forma que $P_t \equiv E(y_t|\boldsymbol{X}) = \Phi(\boldsymbol{X}_t\boldsymbol{\beta})$. Por ser uma função cumulativa de distribuição, $\Phi(x)$ satisfaz automaticamente às

condições (2.15) e (2.16) e, embora não haja fórmula fechada, $\phi(x)$ é facilmente calculada numericamente.

O modelo logit, também nomeado Regressão Logística (RL), é muito semelhante ao modelo probit, possuindo algumas características que o tornam ainda mais simples – a iniciar pela especificação de $F(x)$ como a função logística, $\Lambda(x)$, que possui fórmula fechada.

A modelagem de P_t pela RL é dada por:

$$\Lambda(x) \equiv (1 + e^{-x})^{-1} = \frac{e^x}{1 + e^x} \tag{2.17}$$

$$P_t = \frac{e^{\boldsymbol{X}_t\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_t\boldsymbol{\beta}}} = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta}) \tag{2.18}$$

Em situações práticas, os modelos probit e logit tendem a mostrar resultados muito semelhantes. A maior disseminação do logit, de acordo com Fernandes (2009), somado às características descritas nos parágrafos acima, faz com que a RL seja o modelo utilizado para classificação binária nesta dissertação.

2.2.1
Estimação por Máxima Verossimilhança

Esta seção é baseada em Davidson & MacKinnon (1993).

A estimação de modelos logit é usualmente feita pelo método de Máxima Verossimilhança (MV).

A ideia básica da estimação por MV é determinar um conjunto de parâmetros $\widehat{\boldsymbol{\beta}}$, tal que a verossimilhança de se ter obtido a amostra da qual se trata no experimento seja maximizada – é semelhante a dizer que a fdp ou fp conjunta para o modelo que está sendo estimado seja avaliada nos valores observados da variável dependente da amostra e tratada como uma função somente dos parâmetros. O máximo dessa função é alcançado por $\widehat{\boldsymbol{\beta}}$, o vetor de estimativas de MV. Para utilizar o princípio da verossimilhança e, por consequência, a estimação por MV, parte-se da hipótese de que é possível obter a fdp ou fp da amostra em

questão – um das razões pelas quais se afirma que a MV parte de pressupostos mais fortes e restritivos que a estimação por MQO.

Para que seja realizada a estimação por MV, supõe-se que y seja gerada por uma determinada distribuição de probabilidade. No caso do modelo logit, propõe-se que $\boldsymbol{y}$ seja um vetor de realizações de variáveis aleatórias dicotômicas com distribuição de Bernoulli e que cada uma das n realizações seja oriunda de uma variável aleatória y_t com a mesma distribuição. (2.19) é a fdp para uma variável aleatória qualquer dentro do conjunto citado.

$$f(y_t = j, \boldsymbol{\beta}) = [\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{y_t}[1 - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{1-y_t} \tag{2.19}$$

onde $j \in [0, 1]$ indica possíveis saídas para y_t. Para a RL, supõe-se que $h(\boldsymbol{X}_t, \boldsymbol{\beta}) = \boldsymbol{X}_t\boldsymbol{\beta}$. Assim, $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ é a probabilidade de $y_t = 1$, e $1 - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ é a probabilidade de $y_t = 0$.

A fp conjunta para o modelo que está sendo estimado, função somente dos parâmetros, é denominada função de verossimilhança, $L(\boldsymbol{y}, \boldsymbol{\beta})$. Para qualquer $\boldsymbol{\beta}$, $L(\boldsymbol{y}, \boldsymbol{\beta})$ informa o quão provável é observar a amostra $\boldsymbol{y}$. Pelo fato de $\boldsymbol{y}$ ter sido oriundo de uma AA, a fp conjunta é o produtório da fp de cada y_t.

$$L(\boldsymbol{y}, \boldsymbol{\beta}) = \prod_{t=1}^{n} f(y_t = j, \boldsymbol{\beta}) = \prod_{t=1}^{n} [\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{y_t}[1 - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{1-y_t} \tag{2.20}$$

É usual maximizar a função logaritmo de $L(\boldsymbol{y}, \boldsymbol{\beta})$, nomeada $l(\boldsymbol{y}, \boldsymbol{\beta})$, ao invés de $L(\boldsymbol{y}, \boldsymbol{\beta})$, pela razão de $l(\boldsymbol{y}, \boldsymbol{\beta})$ envolver somatórios e não produtórios. Por $l(\boldsymbol{y}, \boldsymbol{\beta})$ ser monotonicamente crescente, $\widehat{\boldsymbol{\beta}}$ que maximiza $l(\boldsymbol{y}, \boldsymbol{\beta})$ também maximiza $L(\boldsymbol{y}, \boldsymbol{\beta})$.

$$l(\boldsymbol{y}, \boldsymbol{\beta}) = \log L(\boldsymbol{y}, \boldsymbol{\beta}) = \log\left[\prod_{t=1}^{n} [\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{y_t}[1 - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})]^{1-y_t}\right] \tag{2.21}$$

$$l(\boldsymbol{y},\boldsymbol{\beta}) = \sum_{t=1}^{n} [y_t \log\big(\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big) + (1-y_t)\log(1-\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))] \tag{2.22}$$

Esta função é globalmente côncava sempre que $\log\big(\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big)$ e $\log(1-\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))$ forem funções côncavas do argumento $\boldsymbol{X}_t$ – esta condição é satisfeita pelo modelo logit e faz com que as condições de otimização de 1ª ordem tenham solução única. Entretanto, Pratt (1981) não garante que (2.22) seja globalmente côncava quando elementos não lineares, como, por exemplo, termos de iteração entre os regressores, sejam adicionados no modelo. Altman, Gill & McDonald (2003) afirmam que, se a função de maximização é globalmente côncava, os ótimos local e global coincidem.

As condições de 1ª ordem para a maximização de $l(\boldsymbol{y},\boldsymbol{\beta})$ são dadas a seguir. $\widehat{\boldsymbol{\beta}}$ é o estimador de MV que soluciona (2.22).

$$\sum_{t=1}^{n}\left[\frac{(y_t-\Lambda(\boldsymbol{X}_t\widehat{\boldsymbol{\beta}}))\dfrac{\partial\left(\Lambda(\boldsymbol{X}_t\widehat{\boldsymbol{\beta}})\right)}{\partial\widehat{\boldsymbol{\beta}}}X_{ti}}{\Lambda(\boldsymbol{X}_t\widehat{\boldsymbol{\beta}})(1-\Lambda(\boldsymbol{X}_t\widehat{\boldsymbol{\beta}}))}\right] = \mathbf{0} \tag{2.23}$$

Uma forma simplificada é:

$$\sum_{t=1}^{n}(y_t-\Lambda(\boldsymbol{X}_t\widehat{\boldsymbol{\beta}}))X_{ti} \tag{2.24}$$

Como (2.23) não é linear em $\widehat{\boldsymbol{\beta}}$, ela será maximizada numericamente pelo Método de Newton.

2.2.2
Método de Newton

Como $\boldsymbol{y}$ é um vetor de realizações de variáveis aleatórias, $l(\boldsymbol{y},\boldsymbol{\beta})$ é função somente de $\boldsymbol{\beta}$. Logo, é possível reescrever $l(\boldsymbol{y},\boldsymbol{\beta})$ como $l(\boldsymbol{\beta})$.

Supondo-se $\boldsymbol{\beta}^{(0)}$ um palpite inicial para $\widehat{\boldsymbol{\beta}}$, $l(\boldsymbol{\beta})$ uma função duas vezes diferenciável em uma vizinhança em torno de $\boldsymbol{\beta}^{(0)}$ de tal forma que $\boldsymbol{\beta}$ pertença à vizinhança, então, pelo Teorema de Taylor (Graves, 1927):

$$l(\boldsymbol{\beta}) \approx l\left(\boldsymbol{\beta}^{(0)}\right) + \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\right)^{\mathrm{T}} \nabla l\left(\boldsymbol{\beta}^{(0)}\right) + \frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\right)^{\mathrm{T}} \boldsymbol{H}\left(\boldsymbol{\beta}^{(0)}\right)\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\right) = A(\boldsymbol{\beta}) \tag{2.25}$$

$A(\boldsymbol{\beta})$ é a expansão de Taylor de 2ª ordem para $l(\boldsymbol{\beta})$. $\nabla l\left(\boldsymbol{\beta}^{(0)}\right)$ e $\boldsymbol{H}\left(\boldsymbol{\beta}^{(0)}\right)$ são, respectivamente, o vetor gradiente e a matriz hessiana associados à $l(\boldsymbol{\beta})$ em $\boldsymbol{\beta}^{(0)}$. (2.25) é válida para o modelo logit, pois $l(\boldsymbol{\beta})$ é duas diferenciável, por construção do modelo.

O Método de Newton (MN) recebe as informações associadas a $l(\boldsymbol{\beta})$ em $\boldsymbol{\beta}^{(0)}$ (valor da função, gradiente e hessiana) e otimiza $A(\boldsymbol{\beta})$. Como $\boldsymbol{H}\left(\boldsymbol{\beta}^{(0)}\right)$ é negativa definida, a solução pelo MN levará a um máximo, que será global sob as condições citadas anteriormente para ótimo único de $l(\boldsymbol{\beta})$. Diferencia-se $A(\boldsymbol{\beta})$ com relação à $\boldsymbol{\beta}$ e iguala-se a zero para obter:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \boldsymbol{H}^{-1}\left(\boldsymbol{\beta}^{(0)}\right)\nabla l\left(\boldsymbol{\beta}^{(0)}\right) \tag{2.26}$$

O algoritmo continua para o cálculo de $\boldsymbol{\beta}^{(2)}$ e assim por diante. Mesmo que não haja máximo global para $l(\boldsymbol{\beta})$, é garantido que o MN promove valores mais altos para $l(\boldsymbol{\beta})$ (Murray, 2010). Abaixo, a generalização de (2.26):

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} - \boldsymbol{H}^{-1}\left(\boldsymbol{\beta}^{(s)}\right)\nabla l\left(\boldsymbol{\beta}^{(s)}\right) \tag{2.27}$$

$\widehat{\boldsymbol{\beta}}$ é $\boldsymbol{\beta}^{(s+1)}$ que converge em $s+1$ passos para o máximo global ou o valor de $\boldsymbol{\beta}^{(s+1)}$ que faz $l(\boldsymbol{\beta})$ ter maior valor possível.

A implementação utilizada nesta dissertação para obtenção de $\widehat{\boldsymbol{\beta}}$ por MV é uma generalização do MN para o sistema de equações em (2.24) (Deuflhard, 1974).

2.3
Testes de Hipóteses

2.3.1
Definições e Conceitos

De acordo com Davidson & MacKinnon (2003), após a estimação de um modelo, seja linear ou não linear, frequentemente deseja-se testar hipóteses sobre esse modelo. Tais hipóteses geralmente têm a forma de restrições de igualdade ou desigualdade sobre parâmetros do vetor $\boldsymbol{\beta}$. As hipóteses são construídas sobre $\boldsymbol{\beta}$, e não sobre $\widehat{\boldsymbol{\beta}}$, pois a essência do Teste de Hipóteses (TH) é avaliar assertivas sobre o parâmetro populacional ($\boldsymbol{\beta}$) em função de evidência empírica fornecida pela amostra ($\widehat{\boldsymbol{\beta}}$).

A maneira clássica de se realizar um TH é inicialmente formular a hipótese que se deseja testar, definir uma estatística de teste adequada e, em seguida, propor um critério de decisão para rejeitar ou não rejeitar a hipótese, sendo conveniente estruturar este processo em tópicos, como abaixo.

2.3.1.1
Hipóteses Nula e Alternativa

Hipótese, por definição, é uma conjectura sobre a relação entre a variável dependente e um ou mais regressores, expressa através de valores numéricos para $\boldsymbol{\beta}$. Em um TH, interessa-se por testar uma conjectura e, usualmente, inicia-se o TH estipulando-se que conjectura é esta. Por exemplo, pode-se formular uma hipótese onde se deseja testar a conjectura de que algum $\beta_i = 0$, se conjuntamente todos os β_i são iguais à zero ou se $\beta_2 = 3\beta_1$. É comum que esta conjectura seja simples, representando o *status quo* (situação atual) ou caso base de $\boldsymbol{\beta}$ ou algum β_i – dá-se o nome de Hipótese Nula ou $\boldsymbol{H}_0$ a esta conjectura.

Nesta dissertação, $\boldsymbol{H}_0$ para realização de TH em $\boldsymbol{\beta}$ é definida como a seguir:

$$\boldsymbol{H}_0: \beta_i = 0 \tag{2.28}$$

A Hipótese Alternativa, ou $\boldsymbol{H}_1$, é o complemento de $\boldsymbol{H}_0$. Ou seja, quando não há evidência empírica suficiente de que $\boldsymbol{H}_0$ seja verdadeira, rejeita-se $\boldsymbol{H}_0$ e automaticamente não se rejeita $\boldsymbol{H}_1$. Em um TH, somente $\boldsymbol{H}_0$ está sob teste. Rejeitar $\boldsymbol{H}_0$ não gera obrigação de aceitar $\boldsymbol{H}_1$.

As hipóteses $\boldsymbol{H}_0$ e $\boldsymbol{H}_1$ retratadas conjuntamente são dadas por:

$$\begin{aligned} \boldsymbol{H}_0&: \beta_i = 0 \\ \boldsymbol{H}_1&: \beta_i \neq 0 \end{aligned} \tag{2.29}$$

Tanto $\boldsymbol{H}_0$ quanto $\boldsymbol{H}_1$ são conjecturas construídas sobre β_i, um parâmetro populacional. Ao se utilizar uma AA, que é uma porção da população, realiza-se uma inferência sobre o valor numérico populacional de β_i – do qual não se tem conhecimento – a partir do valor numérico estimado $\widehat{\beta_i}$.

2.3.1.2
Estatísticas de Teste

Uma estatística de teste T é uma variável aleatória com fdp ou fp conhecida sob $\boldsymbol{H}_0$. De acordo com a fdp ou fp, calcula-se a probabilidade de T ter sido observado. Se a realização de T é um número real que pode ter ocorrido de maneira aleatória, então não há evidência contra $\boldsymbol{H}_0$. Caso o contrário ocorra, há evidência contra $\boldsymbol{H}_0$ e é necessário um critério de decisão para rejeitá-la.

Em função de T, determina-se outra estatística de teste denominada p-valor que, por definição, é o menor nível de significância no qual $\boldsymbol{H}_0$ é rejeitada.

2.3.1.3
Critério de Decisão

O critério de decisão é função do nível de significância do teste. Nível de significância ou tamanho do teste é a probabilidade de que T rejeite $\boldsymbol{H}_0$ quando

$\boldsymbol{H}_0$ é verdadeira. Sendo β_i o parâmetro que se deseja testar e $\ominus_0 = 0$ o conjunto de valores que satisfaz $\boldsymbol{H}_0$, define-se o tamanho do teste, α, de $\boldsymbol{H}_0$:

$$\alpha \equiv \Pr(T \in RR | \beta_i = 0) \tag{2.30}$$

Se T for usada como estatística de teste, define-se um critério de decisão através da divisão da região de possíveis valores de T em duas sub-regiões, denominadas Região de Rejeição (RR) de $\boldsymbol{H}_0$ e Região de Não Rejeição (RNR) de $\boldsymbol{H}_0$: se T assumir valor dentro de RNR, $\boldsymbol{H}_0$ não é rejeitada; se T assumir valor dentro de RR, $\boldsymbol{H}_0$ é rejeitada.

Se o p-valor, função de T, for usado como estatística de teste, o critério de decisão passa a ser não rejeitar $\boldsymbol{H}_0$ se p-valor $\geq \alpha$ e rejeitar $\boldsymbol{H}_0$ se p-valor $< \alpha$.

2.3.2
TH em Modelos de Regressão Linear

Até esta seção, nada foi dito com relação às propriedades estatísticas da estimação por MQO. No início do capítulo, foi feita referência à distinção entre os conjuntos de propriedade numérica e estatística – a última tem como base hipóteses sobre a forma como o conjunto de dados foi gerado.

Davidson & MacKinnon (1993) afirmam que o erro $\boldsymbol{u}$, em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ (2.1), é a única forma pela qual a aleatoriedade é inserida na variável dependente de um modelo de regressão. Conclui-se, portanto, que a aleatoriedade de $\boldsymbol{y}$ em (2.1) necessariamente virá de $\boldsymbol{u}$ e que hipóteses serão necessárias para sustentar a sua natureza aleatória. Tais hipóteses são as listadas em seguida.

Hipótese 1: $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{2.31}$$

Hipótese 2: $$E(\boldsymbol{u}|\boldsymbol{X}) = \boldsymbol{0} \tag{2.32}$$

Hipótese 3: $$\rho(\boldsymbol{X}) = k \tag{2.33}$$

Hipótese 4: $$Var\,(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2 \boldsymbol{I}_n \tag{2.34}$$

Hipótese 5: $$u_t|\boldsymbol{X} \sim \text{NID}(0, \sigma^2) \quad \therefore \quad \boldsymbol{u}|\boldsymbol{X} \sim \text{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \tag{2.35}$$

A Hipótese 1 requer que o modelo seja linear nos parâmetros, ou seja, possa ser escrito sob a forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. A Hipótese 2 é denominada hipótese de média condicional nula, que implica $Corr(\boldsymbol{u}, \boldsymbol{X}) = 0$ e, segundo Wooldridge (2008), é sugerida caso o conjunto de dados seja amostrado aleatoriamente. A Hipótese 3 requer inexistência de colinearidade perfeita. A Hipótese 4 diz respeito à necessidade de homocedasticidade (variância constante) e ausência de correlação serial entre os erros das observações – esta última é plenamente atendida sob amostragem aleatória ou qualquer outro processo de amostragem de corte transversal com observações independentes. Na Hipótese 5 determina-se a necessidade de normalidade de $\boldsymbol{u}|\boldsymbol{X}$.

Sob as Hipóteses 1 à 5 e utilizando-se da teoria de combinação de variáveis aleatórias (Casella & Berger, 2011), sendo $\boldsymbol{u} = [u_1\ u_2\ \cdots\ u_n]^{\text{T}}$ um vetor de variáveis aleatórias e supondo-se (2.35), determina-se que $\boldsymbol{y}|\boldsymbol{X}$ também será vetor de variáveis aleatórias tal como (2.36), fazendo com que $\widehat{\boldsymbol{\beta}}$ também o seja – como mostra (2.37).

$$\boldsymbol{y}|\boldsymbol{X} \sim \text{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) \tag{2.36}$$

$$\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}) \tag{2.37}$$

onde $Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$ em (2.37) e $\boldsymbol{I}_n$ é a matriz identidade n x n. (2.37) é considerada a base da inferência estatística envolvendo $\boldsymbol{\beta}$ (Wooldridge, 2006).

σ^2 é um parâmetro populacional e, portanto, desconhecido. Seu estimador não tendencioso é dado por:

$$\hat{\sigma}^2 = \frac{\widehat{\boldsymbol{u}}^{\mathrm{T}}\widehat{\boldsymbol{u}}}{n-k-1} \tag{2.38}$$

Quando, em (2.37), substitui-se σ^2 por $\hat{\sigma}^2$, a grandeza T assume uma distribuição *t-student* com $n-k-1$ graus de liberdade (Wooldridge, 2006), sendo $SE(\hat{\beta}_i) = \sqrt{Var(\hat{\beta}_i)}$.

$$T = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)/\sqrt{n}} \sim t\,(n-k-1) \tag{2.39}$$

Com as informações sobre a distribuição de T, é possível caracterizar completamente o TH para $\boldsymbol{\beta}$ em modelos de regressão linear.

2.3.2.1
Hipóteses Nula e Alternativa

Seguem o modelo de (2.29).

2.3.2.2
Estatísticas de Teste e Critério de Decisão

T, em (2.39), é uma estatística de teste pois é uma variável aleatória com fdp conhecida sob $\boldsymbol{H}_0$ e somente assume a distribuição em (2.39) devido à

distribuição condicional de $\widehat{\boldsymbol{\beta}}$ em (2.37). A estatística de teste T tem fdp como mostrado na Figura 2.3.

Nesta dissertação, embora o separador decimal seja a vírgula e o separador de milhar seja o ponto, algumas Figuras apresentarão o ponto como separador decimal, como na Figura 2.3, pelo fato destas Figuras terem sido originadas de *softwares* originalmente desenvolvidos em países de língua inglesa, que utilizam o ponto como separador decimal. Somente algumas Figuras (e não tabelas) apresentam essa troca, o que não deve gerar problemas de entendimento.

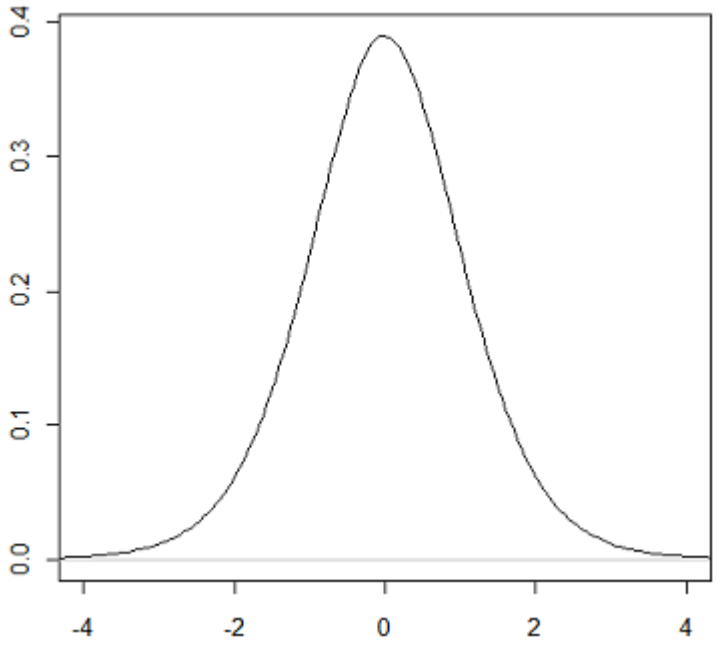

Figura 2.3 – fdp de $\boldsymbol{T}$

Fonte: adaptado de Wooldridge (2006)

Atribuídos n e k à (2.39), T está plenamente definida, pois $n-k-1$ é parâmetro único da distribuição. Sendo T uma variável aleatória, pode-se calcular probabilidades ou outras grandezas associadas à T. Por exemplo, pode-se determinar os quantis $-t_{\alpha/2}(n-k-1)$ e $t_{\alpha/2}(n-k-1)$ que fazem com que a probabilidade de T estar entre $-t_{\alpha/2}(n-k-1)$ e $t_{\alpha/2}(n-k-1)$ seja $1-\alpha$. Para efeito de simplificação de notação, serão utilizados os termos $-t_{\alpha/2}$ e $t_{\alpha/2}$. A Figura 2.4 identifica as grandezas citadas.

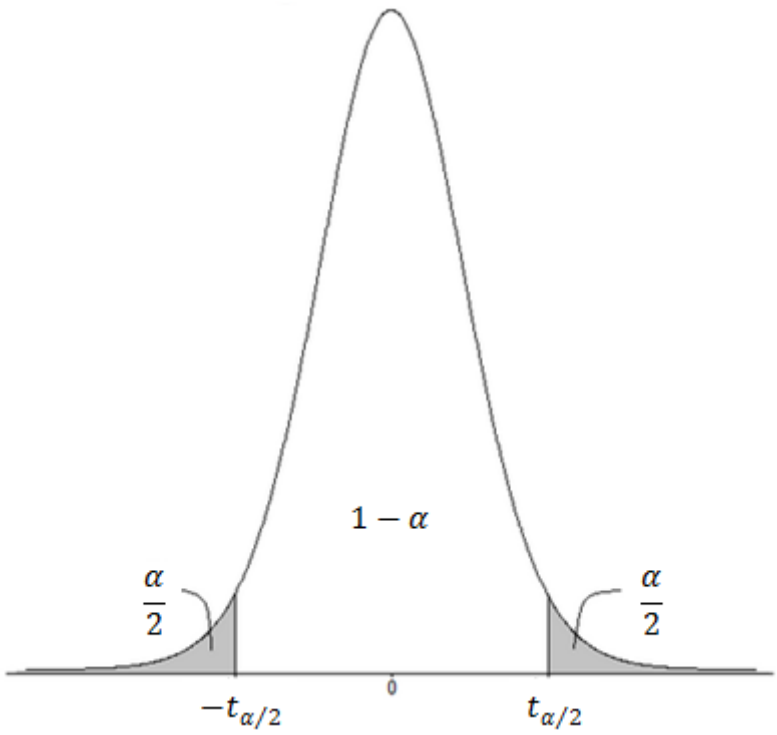

Figura 2.4 – fdp de $\boldsymbol{T}$ e grandezas relativas à $\boldsymbol{T}$

Se T_{obs} – a realização da variável aleatória T quando os valores de $\hat{\beta}_i$, β_i, $SE(\hat{\beta}_i)$ e $\sqrt{n}$ são substituídos em (2.39) – é um valor na região cinza da Figura 2.4, é pouco provável, sob o nível de confiança de $1-\alpha$, que T_{obs} seja de fato um valor oriundo da distribuição de T sob $\boldsymbol{H}_0$. Nesse caso, deve-se rejeitar $\boldsymbol{H}_0$, pois T_{obs} está dentro da RR de $\boldsymbol{H}_0$. É fundamental atentar que β_i é um parâmetro populacional, portanto desconhecido, mas que, ao se realizar o TH, atribui-se a ele, em (2.39), o seu valor sob $\boldsymbol{H}_0$. Tendo (2.29) como referência para TH, β_i receberá o valor 0 para que T_{obs} seja obtido.

Se T_{obs} é um valor na região branca (compreendida entre a curva e o eixo horizontal) da Figura 2.4, é muito provável, sob o nível de confiança de $1-\alpha$, que T_{obs} seja de fato um valor oriundo da distribuição de T sob $\boldsymbol{H}_0$. Nesse caso, deve-se não rejeitar $\boldsymbol{H}_0$, pois T_{obs} está dentro da RNR de $\boldsymbol{H}_0$.

Os quatro parágrafos anteriores, que definem T como estatística de teste e determinam critérios de decisão claros e objetivos relacionados à T, integrados à (2.29), caracterizam um teste bicaudal para β_i. Um TH bicaudal é tal que se considera que desvios do parâmetro estimado ($\hat{\beta}_i$) sejam teoricamente possíveis

em qualquer direção (positiva ou negativa), tendo como referência (2.29), de acordo com Gujarati (2008).

Calcula-se o p-valor para que este possa ser usado como estatística de teste. No caso de TH bilaterais com $\boldsymbol{H}_0$ e $\boldsymbol{H}_1$ como em (2.29), o p-valor é calculado como em (2.40).

$$\text{p-valor} = \Pr(|T| > |T_{obs}|) \tag{2.40}$$

A Figura 2.5 ilustra o p-valor (região total hachurada em vermelho) para o caso de T_{obs} estar na RR.

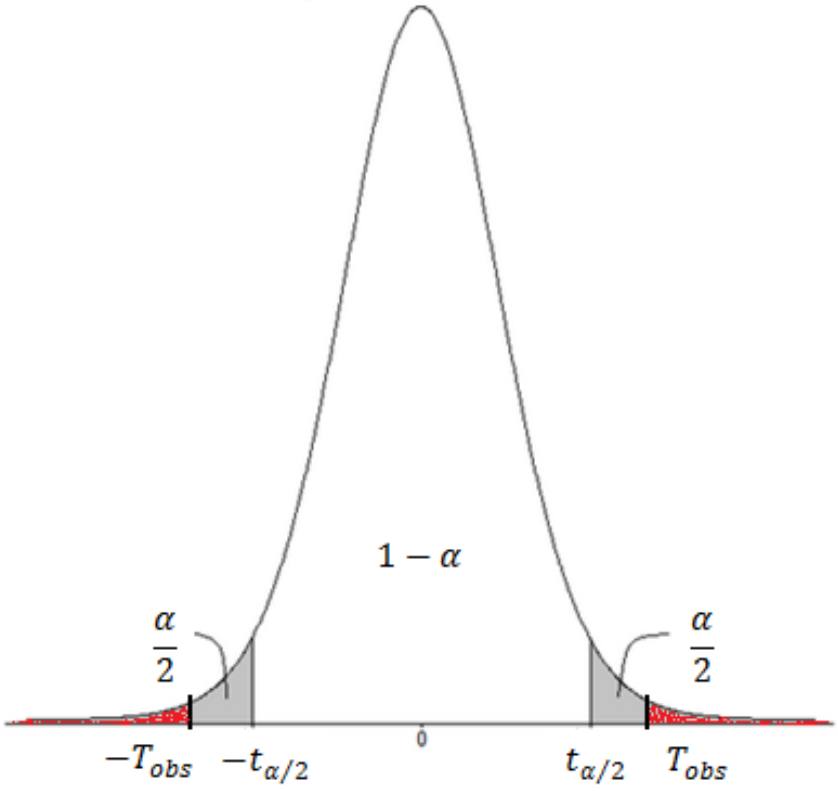

Figura 2.5 – Ilustração do p-valor

Para efeito de exemplificação, supõe-se $\alpha = 5\%$ e p-valor $= 1\%$. Com as evidências numéricas $-T_{obs}$ e T_{obs}, supostamente advindas de T sob $\boldsymbol{H}_0$, observa-se que 1% é o menor nível de significância com que se rejeita $\boldsymbol{H}_0$. Como $5\% > 1\%$, $\boldsymbol{H}_0$ será rejeitada sob o valor de α estipulado – para todos os valores

de α maiores que 1%, $\boldsymbol{H}_0$ será rejeitada. Nesta dissertação, utiliza-se o p-valor como estatística de teste para testes de hipóteses.

Pode-se escrever (2.39) como:

$$T_{n_{gl}} = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)/\sqrt{n}} \sim t\ (n_{gl}) \tag{2.41}$$

Considera-se, para efeito de simplificação de notação, que $n_{gl} = n - k - 1$ (n_{gl} é o número de graus de liberdade). Variando-se n_{gl} sequenciamente, $\{T_{n_{gl}}\}$ torna-se uma sequência de variáveis aleatórias, todas com distribuição *t-student*. A fdp de $T_{n_{gl}}$, para algum n_{gl}, é dada por:

$$f_{n_{gl}}(t_{\%}) = \frac{\Gamma\left(\frac{n_{gl}+1}{2}\right)}{\Gamma\left(\frac{n_{gl}}{2}\right)\sqrt{\pi n_{gl}}}\left(1 + \frac{{t_{\%}}^2}{n_{gl}}\right)^{\frac{-(n_{gl}+1)}{2}} \tag{2.42}$$

$\Gamma(\cdots)$ é a função gamma, definida por $\Gamma(n_{gl}) = (n_{gl} - 1)!$, uma extensão da função fatorial. Embora $\Gamma(\cdots)$ seja aplicada a números reais e complexos, a definição anterior se aplica a n_{gl} inteiros e positivos. A expressão $t_{\%}$ é o quantil associado a algum percentil da distribuição – por isso a notação "$t_{\%}$".

Conforme $n_{gl} \to \infty$, é natural buscar entender como se comporta $f_{n_{gl}}(t_{\%})$. Logo, deseja-se calcular:

$$\begin{aligned}\lim_{n_{gl}\to\infty} f_{n_{gl}}(t_{\%}) &= \lim_{n_{gl}\to\infty}\left[\frac{\Gamma\left(\frac{n_{gl}+1}{2}\right)}{\Gamma\left(\frac{n_{gl}}{2}\right)\sqrt{\pi n_{gl}}}\left(1 + \frac{{t_{\%}}^2}{n_{gl}}\right)^{\frac{-(n_{gl}+1)}{2}}\right] \\ &= \lim_{n_{gl}\to\infty}\left[\frac{\Gamma\left(\frac{n_{gl}+1}{2}\right)}{\Gamma\left(\frac{n_{gl}}{2}\right)\sqrt{\pi n_{gl}}}\right]\lim_{n_{gl}\to\infty}\left[\left(1 + \frac{{t_{\%}}^2}{n_{gl}}\right)^{\frac{-(n_{gl}+1)}{2}}\right]\end{aligned} \tag{2.43}$$

Utilizando-se a aproximação de Stirling (Marsaglia & Marsaglia, 1990), (2.43) pode ser escrita como:

$$\lim_{n_{gl}\to\infty} f_{n_{gl}}(t_{\%}) = \lim_{n_{gl}\to\infty}\left[\frac{\Gamma\left(\frac{n_{gl}+1}{2}\right)}{\Gamma\left(\frac{n_{gl}}{2}\right)\sqrt{\pi n_{gl}}}\right]\lim_{n_{gl}\to\infty}\left[\left(1+\frac{{t_{\%}}^2}{n_{gl}}\right)^{\frac{-(n_{gl}+1)}{2}}\right] \quad (2.44)$$

$$= \left[\frac{1}{\sqrt{2\pi}}\right]\left[\exp\left(-\frac{{t_{\%}}^2}{2}\right)\right]$$

Esta equação mostra que o $\lim_{n_{gl}\to\infty} f_{n_{gl}}(t_{\%})$ é a fdp da distribuição normal padronizada. Portanto, $T_{n_{gl}} \xrightarrow{d} N(0,1)$ e, para n_{gl} suficientemente grande, os modelos de TH discutidos até então permanecem válidos, somente devendo-se reconhecer que não será mais a distribuição *t-student* a utilizada – e sim a normal padrão, já que a 1ª converge em distribuição para a 2ª. Como $n \gg k$, para os modelos desenvolvidos nesta dissertação, n_{gl} será suficientemente grande para que haja a convergência em distribuição. Casella & Berger (2011) fornecem um tratamento adequado para tipos de convergência.

Para efeito de ilustração, propõe-se a Figura (2.6). A curva em verde mais claro é a fdp de uma variável aleatória com distribuição *t-student*, $n_{gl} = 1$. As curvas de cor verde escurecem conforme n_{gl} aumenta. Quanto maior n_{gl}, mais semelhante à curva da fdp normal padrão, em amarelo, as curvas de *t-student* se tornam. Para $n_{gl} = 30$ (n_{gl} está representado pela sigla "d.f.", somente nesta Figura), as curvas da *t-student* e normal padrão praticamente se sobrepõem.

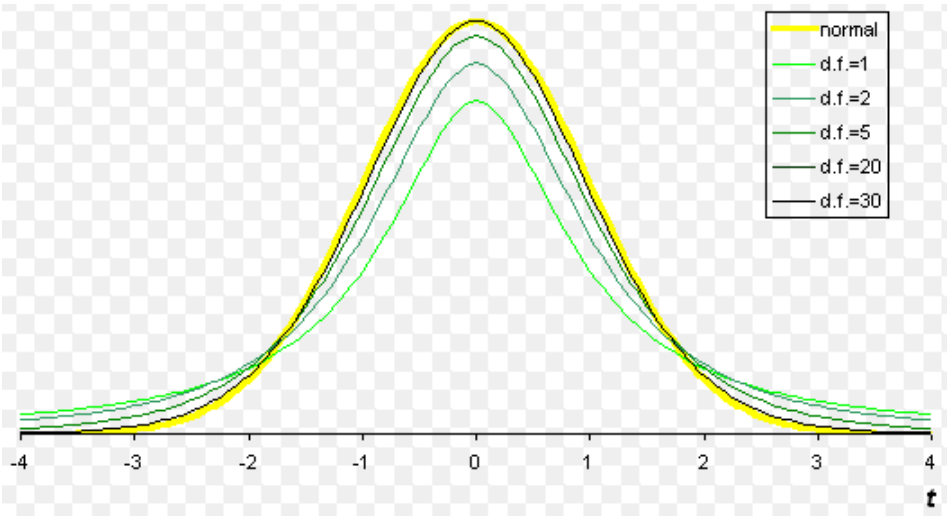

Figura 2.6 – Aproximação de $T_{n_{gl}}$ por $N(0,1)$

2.3.3
TH em Modelos de Regressão Não Linear: Classificação Binária

De acordo com Davidson & MacKinnon (1993), os modelos logit satisfazem as três condições de regularidade necessárias para que $\widehat{\boldsymbol{\beta}}$, estimado por MV na seção 2.2.1, seja uma variável aleatória que tenha distribuição assintoticamente normal, com matriz de variância-covariância dada pela inversa da matriz de informação.

Segundo Greene (2011), a primeira condição requer que as diferenciais em (2.45) existam, sejam funções contínuas e finitas para quase todos os elementos de $\boldsymbol{y}$ e para todos os elementos de $\boldsymbol{\beta}$ que pertencem a um intervalo não degenerativo (intervalo que possui comprimento estritamente positivo). Esta condição garante a existência de uma aproximação por série de Taylor e de variâncias finitas das diferenciais de $l(\boldsymbol{y}, \boldsymbol{\beta})$.

$$\frac{\partial}{\partial \boldsymbol{\beta}}\big(l(\boldsymbol{y}, \boldsymbol{\beta})\big), \frac{\partial^2}{\partial \boldsymbol{\beta}^2}\big(l(\boldsymbol{y}, \boldsymbol{\beta})\big) \; e \; \frac{\partial^3}{\partial \boldsymbol{\beta}^3}\big(l(\boldsymbol{y}, \boldsymbol{\beta})\big) \tag{2.45}$$

A segunda condição define que os valores esperados das duas primeiras diferenciais de $l(\boldsymbol{y}, \boldsymbol{\beta})$ possam ser calculados. Como $\partial/\partial \boldsymbol{\beta}\,\big(l(\boldsymbol{y}, \boldsymbol{\beta})\big)$ e $\partial^2/\partial \boldsymbol{\beta}^2\,\big(l(\boldsymbol{y}, \boldsymbol{\beta})\big)$ são funções aleatórias, espera-se que cumpram com os requisitos tradicionais para cálculos de valores esperados – tais requisitos estão em Casella & Berger (2011).

A terceira condição requer, para todo $\boldsymbol{\beta}$ pertencente ao intervalo não generativo citado anteriormente, que a terceira diferencial de $l(\boldsymbol{y}, \boldsymbol{\beta})$ em relação a elementos distintos de $\boldsymbol{\beta}$ seja menor que uma função que possui valor esperado finito. Esta condição fará com que a série de Taylor proposta anteriormente seja truncada.

Sob as condições citadas e n suficientemente grande (o menor n, para todos os conjuntos citados nesta dissertação, é 200):

$$\widehat{\boldsymbol{\beta}} \xrightarrow{d} N\left(\boldsymbol{\beta}, \boldsymbol{\sigma}_{\boldsymbol{\beta}}^2\right) \tag{2.46}$$

Onde $\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2$ é chamada de variância assintótica de $\widehat{\boldsymbol{\beta}}$. Greene (2011) prova que $\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2 = [I(\boldsymbol{\beta})]^{-1}$, sendo $I(\boldsymbol{\beta})$ a informação de Fisher:

$$I(\boldsymbol{\beta}) = -E_{\boldsymbol{\beta}}\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2} l(\boldsymbol{y}, \boldsymbol{\beta})\right] \tag{2.47}$$

$I(\boldsymbol{\beta})$ deve ser estimada. A forma de cálculo em (2.47) raramente estará disponível, segundo Greene (2011), que propõe duas alternativas ao seu cálculo: (1) a matriz atual (e não o valor esperado) de derivadas segundas de $l(\boldsymbol{y}, \boldsymbol{\beta})$ na estimativa de MV e (2) a matriz de variâncias-covariâncias do vetor gradiente de $l(\boldsymbol{y}, \boldsymbol{\beta})$. A diferença entre os estimadores está no custo computacional em certas situações, mas todos são assintoticamente equivalentes.

(2.46) permite que se faça TH para RL de forma semelhante ao TH que se faz para modelos de regressão linear. O parâmetro β_i terá distribuição sob $\boldsymbol{H}_0$ como em (2.48), quando n é suficientemente grande:

$$U = \frac{\hat{\beta}_i - \beta_i}{SE\left(\hat{\beta}_i\right)} \sim N(0,1) \tag{2.48}$$

sendo U a estatística de teste. A estatística de teste T converge em distribuição para $N(0,1)$, como já visto. $SE\left(\hat{\beta}_i\right)$ é a raiz quadrada do elemento correspondente ao regressor i da diagonal principal de $\boldsymbol{\sigma}_{\boldsymbol{\beta}}^2$.

É comum que se realize TH para β_i na RL utilizando-se o quadrado de U. Nesse caso, $U^2 \sim \chi^2(1)$. Pelo fato do quadrado de uma variável aleatória normal padrão ter distribuição $\chi^2(1)$, U e U^2 levam ao mesmo resultado quando se aplica um TH nos moldes de (2.29).

O TH para modelos de RL está definido: segue-se o modelo proposto em (2.29) para $\boldsymbol{H}_0$ e $\boldsymbol{H}_1$; U foi proposta como estatística de teste mas será o p-valor, função de U, a referência para tomada de decisão e continuará a ser utilizado como estatística de teste. O cálculo do p-valor é semelhante ao apresentado anteriormente para modelos de regressão linear.

3
Programação Genética

A PG é a ferramenta que realiza o processo de evolução e geração de modelos desta dissertação – este capítulo apresenta seus fundamentos.

O conteúdo específico sobre PG deste capítulo é baseado em Poli et al. (2008). A não ser quando se explicita o contrário, a sigla PG se refere à PG tradicional, proposta por Koza (1992).

3.1
Introdução

Em ciência da computação, a computação evolucionária é um sub-ramo da inteligência computacional. Esta compreende um conjunto de metodologias computacionais e abordagens inspiradas em elementos da natureza, como a seleção natural, para resolver problemas tais como as tarefas de regressão e classificação aqui abordadas.

A PG é uma técnica sistemática da computação evolucionária que automaticamente resolve problemas sem a necessidade de se conhecerem informações do domínio do conjunto de dados, do problema ou formato da solução do problema.

Em PG, evolui-se uma população de programas de computador: geração a geração, a PG estocasticamente transforma uma população de programas em uma nova população de programas – espera-se que os novos programas sejam melhores do que aqueles que os geraram, embora tal resultado não possa ser garantido, mesmo com ferramentas de elitismo. O elitismo copia o melhor indivíduo da população p para a população $p + 1$, de acordo com uma métrica de aptidão. A acurácia representa a aptidão – conceito que define a qualidade de uma solução ou programa – para tarefas de regressão e classificação.

A PG, por sua natureza estocástica e disponibilidade de operadores de mutação e cruzamento, tem a capacidade de explorar com abrangência o espaço de busca do problema, evitando a permanência em extremos locais. Mutação e

cruzamento são operadores utilizados para criar novos programas a partir de outros já existentes: enquanto o primeiro operador cria um novo programa (também chamado de indivíduo ou modelo) a partir de uma alteração aleatória em uma parte do programa original (esta parte também é escolhida aleatoriamente), o segundo cria um novo programa a partir da combinação de partes de dois programas (tanto os programas quanto as partes dos programas que serão combinadas são escolhidas aleatoriamente). Mutação e cruzamento são chamados de operadores genéticos.

A Figura 3.1 apresenta o pseudocódigo genérico de um algoritmo de PG.

```
1: Crie aleatoriamente uma população inicial de programas,
   a partir das primitivas disponíveis.
2: REPITA
3:     Execute cada programa e determine sua acurácia.
4:     Selecione um ou dois programas da população, em função
       de probabilidades determinadas pela acurácia, para que
       façam parte do processo de criação da nova população,
       utilizando os operadores de mutação ou cruzamento.
5:     Crie novos programas, aplicando os operadores citados,
       de acordo com probabilidades do experimento (são distintas
       das probabilidades determinadas pela acurácia).
6: ATÉ QUE: uma solução aceitável seja encontrada ou outra
   condição de parada seja atingida (por exemplo, o número
   máximo de gerações).
7: RETORNE: o indivíduo de melhor acurácia.
```

Figura 3.1 – Pseudocódigo genérico de um algoritmo de PG

Fonte: adaptado de Poli et al. (2008)

Este pseudocódigo servirá de base para a descrição da PG.

3.2 Representação

A representação dos programas é uma importante escolha a ser tomada em um experimento de PG – tal decisão antecede o 1º passo do algoritmo de PG, visto na Figura 3.1.

Em PG, os programas são usualmente representados por árvores, diferindo da representação por linhas de código. A Figura 3.2 ilustra a representação do

programa $\max(x + x, x + 3 * y)$. As variáveis e constantes do programa (x, y e 3) são os elementos extremos da árvore, nomeados folhas ou terminais. As operações aritméticas ($+, *$ e max) são nós internos, denominados funções. O conjunto de terminais e funções é denominado conjunto de primitivas de um experimento de PG, referenciado por ϑ. O conjunto de terminais (variáveis e constantes) será referenciado por Ω nesta dissertação.

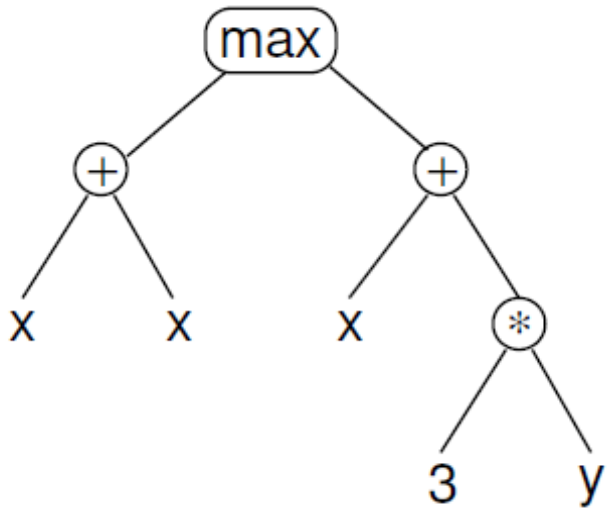

Figura 3.2 – Representação em árvore do programa $\mathbf{max}(\boldsymbol{x} + \boldsymbol{x}, \boldsymbol{x} + \mathbf{3} * \boldsymbol{y})$

Fonte: Poli et al. (2008)

A árvore da Figura 3.2 pode ser um programa de um experimento de PG ou parte de um programa. Se é parte de um programa, dá-se o nome de gene, ramo ou sub-árvore – pode inclusive ser chamada de componente ou elemento. Um programa pode ser estruturado de uma forma mais complexa, composto por um conjunto de genes, como mostra a Figura 3.3 – esta estrutura de representação é chamada de multigênica. A utilização de um nó especial, denominado nó raiz, unifica os genes em uma árvore única.

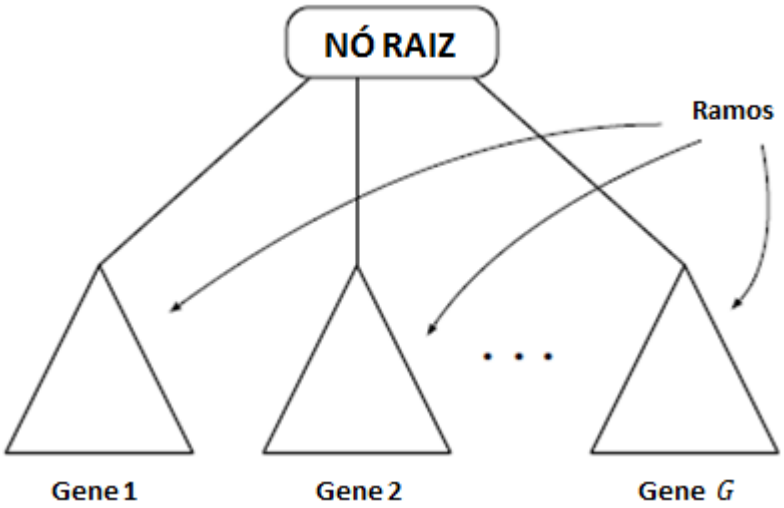

Figura 3.3 – Representação multigênica de um indivíduo de PG

Fonte: adaptado de Poli et al. (2008)

Na PG, não é incomum que o conjunto de primitivas seja composto, além das funções, por constantes denominadas constantes efêmeras. Usualmente, são geradas aleatoriamente dentro de um intervalo real.

3.3
1º Passo: Criação da População Inicial

Como em outros algoritmos evolucionários, os indivíduos da população inicial da PG são tipicamente gerados de maneira aleatória – há alguns métodos que desempenham essa tarefa, tal como o método *full*, método *grow* e método *hamped half-and-half*. As metodologias citadas utilizarão ϑ para gerar os indivíduos.

Tanto no método *full* quanto no método *grow*, os indivíduos são gerados de tal forma que não ultrapassem uma altura máxima. A altura de um nó é o número de arestas que precisam ser percorridas para que se acesse o nó em questão, partindo do nó raiz, que, assume-se, tem altura igual à zero. A altura de uma árvore é a altura do nó/folha mais distante do nó raiz.

Na metodologia *full* – assim chamada pelo fato de gerar árvores cheias, ou seja, todos os nós estão a uma mesma altura – os nós são preenchidos aleatoriamente com funções oriundas do conjunto de funções, até que a altura

máxima da árvore seja atingida. Após o preenchimento, por funções, de todos os nós até a altura da árvore, terminais são aleatoriamente selecionados de Ω para compor a árvore no nível seguinte à altura da árvore. Embora o método *full* crie programas com folhas contendo sempre a mesma altura, não necessariamente todos os indivíduos da população inicial terão um número de nós (grandeza referenciada como tamanho de uma árvore na PG) idêntico em sua estrutura e/ou o mesmo formato – as duas situações só acontecerão se todas as funções em ϑ receberem o mesmo número de argumentos. Caso recebam, os programas gerados na população inicial, em função de tamanho e formato, tendem a ser muito parecidos.

O método *grow* permite a inicialização de árvores de tamanhos e formatos mais variados. Os nós são selecionados aleatoriamente de ϑ, até que a altura da árvore seja atingida – qualquer nó da árvore, até atingir sua altura, pode ser preenchido por uma função, variável ou constante, diferentemente do método *full*, que somente permitia preencher com funções os nós até a altura da árvore. Assim como o método *full*, o método *grow* preenche os nós seguintes à altura da árvore somente com terminais.

*Mesmo a metodologia *grow* tendo possibilitado a criação de programas com variabilidade um pouco maior em tamanho e formato, Koza (1992) propôs um terceiro método, *hamped half-and-half*, que combina *full* e *grow*: metade da população inicial é construída utilizando-se o método *full* e, a outra metade, o *grow*. Isso é possível a partir da utilização de um intervalo de alturas limites para árvores, garantindo que árvores com distintos tamanhos e formatos serão geradas.

O tamanho e o formato mais frequentes gerados para os indivíduos da população inicial variarão de acordo com o método utilizado. As três metodologias descritas estão sujeitas à distribuição das variáveis, constantes e funções em ϑ, fazendo com que o valor esperado mais frequente de tamanho e formato de indivíduos seja uma informação pouco tangível, devido também ao aspecto aleatório inerente à formação de indivíduos da PG e seu poder combinatório.

3.4
2º Passo: Estrutura de Repetição

Construída a população inicial, o algoritmo de PG propõe a evolução de seus indivíduos através de uma estrutura de repetição. A população inicial, também chamada de geração inicial, será exposta a rotinas que buscam fazer com que, em média, os indivíduos das populações/gerações subsequentes sejam melhores do que os indivíduos das populações/gerações anteriores, em função de uma métrica de acurácia. Este processo finda quando uma solução com acurácia aceitável é encontrada ou outra condição de parada é atingida.

3.5
3º Passo: Determinação e Cálculo da Acurácia

A determinação e o cálculo da acurácia são realizados somente após a definição do conjunto de terminais, Ω, e do conjunto de funções. A definição de Ω e do conjunto de funções é função do tipo de problema em estudo. O conjunto de terminais Ω pode ser composto de variáveis e constantes efêmeras, além de funções sem argumentos como, por exemplo, a função que gera um número aleatório seguindo uma distribuição uniforme dentro de um intervalo real. O conjunto de funções pode ser composto por funções aritméticas (soma, subtração, multiplicação e divisão), matemáticas (seno, cosseno, tangente, etc.), booleanas (ou, e, não, etc.), condicionais (se-então) ou de repetição.

Para que a PG funcione efetivamente, é comum requisitar que o conjunto de funções de um experimento tenha a propriedade de fechamento (*closure*) – que pode ser entendida como a união de duas outras propriedades, chamadas de consistência de tipo e segurança na avaliação – e a propriedade de suficiência.

A consistência de tipo é requerida devido ao operador de cruzamento poder intercambiar nós arbitrariamente entre os indivíduos participantes da operação. Consequentemente, é necessário que qualquer gene possa ser usado como argumento em qualquer posição de qualquer função do conjunto de funções. Portanto, é comum que todas as funções tenham consistência de tipo, ou seja, todas retornam valores do mesmo tipo, sendo que seus argumentos também possuam este tipo.

O outro componente de *closure* é a segurança na avaliação. Essa propriedade é requerida pelo fato de muitas funções usualmente utilizadas em PG falharem ao rodar em tempo real. Este problema é tipicamente tratado através de uma modificação na natureza das funções do conjunto de primitivas ϑ. Uma modificação tradicional em funções de ϑ é a versão protegida de funções numéricas, que potencialmente podem ocasionar erros em tempo de execução devido ao seu domínio limitado, tais como divisão e operações com logaritmos.

A propriedade de suficiência determina ser possível expressar uma solução à tarefa de interesse usando elementos de ϑ. Mais formalmente, ϑ será dito suficiente se o conjunto de todas as possíveis combinações de elementos de ϑ gerarem ao menos uma solução válida para a tarefa de interesse.

A (métrica de) acurácia ou função objetivo é a grandeza responsável por identificar quais regiões do espaço de busca podem ser determinadas como as mais prováveis de fornecer programas que solucionem, plenamente ou aproximadamente, a tarefa de interesse. O espaço de busca a ser explorado pela PG, definido como todas as possíveis soluções para a tarefa em questão, também é função do conjunto de terminais (Ω) e do conjunto de ϑ.

A acurácia pode ser mensurada de distintas maneiras. Por exemplo, em termos da: quantificação de uma grandeza de erro entre o estimado pela PG e a variável de resposta; quantidade de tempo, combustível ou unidades monetárias necessárias para levar um sistema a um estado estacionário; acurácia do programa em reconhecer padrões ou classificar objetos; determinação do *payoff* que um jogo gera ao jogador; dentre outras maneiras.

3.6
4º Passo: Seleção

O propósito da operação de seleção é escolher indivíduos, entre o total de indivíduos da população, para as operações de cruzamento e mutação, responsáveis pela criação da população seguinte.

Assim como na maior parte dos algoritmos evolucionários, os operadores genéticos em PG são aplicados aos indivíduos, aleatoriamente selecionados, baseando-se na acurácia. Ou seja, indivíduos com maior acurácia tendem a perpetuar os seus genes nas gerações seguintes; enquanto que, para os indivíduos

com menor acurácia, a mesma tendência não ocorre. O método de torneio é o mais comumente usado para seleção de indivíduos em PG.

Na seleção por torneio, inicialmente escolhe-se aleatoriamente um determinado número de indivíduos dentro do total de indivíduos da população ($n_{torneio}$). Os indivíduos são comparados uns com os outros, utilizando a acurácia como métrica de comparação, e somente o vencedor do torneio é escolhido para ser o indivíduo que será utilizado na operação de cruzamento ou mutação – também denominado indivíduo gerador.

Em algoritmos evolucionários, é comum que o indivíduo gerador seja nomeado "genitor", e o indivíduo gerado, a partir do indivíduo gerador, "descendente". Caso haja dois indivíduos geradores, intercambiando genes para gerar um novo, os indivíduos geradores são nomeados "genitores" do indivíduo gerado. Na operação de cruzamento, são necessários dois indivíduos genitores; consequentemente são realizados duas seleções por torneio (uma para cada genitor). Na operação de mutação, só há necessidade de um indivíduo gerador e, por consequência, somente uma seleção por torneio é realizada.

A seleção por torneio somente dá peso a qual dos indivíduos é melhor entre os escolhidos aleatoriamente para participar do torneio; ela não evidencia o quanto o vencedor do torneio é melhor que os outros. Portanto, um indivíduo com acurácia muito elevada frente aos outros de sua população não poderia perpetuar seus genes, através de seus filhos, de maneira maciça na população seguinte. Caso isso acontecesse, uma rápida perda de diversidade ocorreria. Há um método de seleção, denominado seleção por proporcionalidade de acurácia ou seleção por roleta, em que as chances da situação anteriormente citada ocorrer são maiores. Na seleção por torneio, essas chances são pequenas.

Usualmente, $n_{torneio}$ indivíduos são aleatoriamente escolhidos na população – é permitido que um indivíduo seja selecionado mais de uma vez. Caso um ou mais indivíduos tenham a mesma acurácia, é comum que se aplique na PG a pressão lexicográfica proposta por Luke & Panait (2002), que consiste em escolher, dentre os indivíduos de acurácia igual, aquele com o menor número de nós.

3.7
5º Passo: Mutação, Cruzamento e Elitismo

Esta dissertação contemplará as seguintes formas de mutação:

(1) Mutação tradicional proposta por Koza (1992): inicialmente, seleciona-se um indivíduo; em seguida, um gene desse indivíduo; por último, um nó do gene – todas essas escolhas são aleatórias. Deve-se então excluir a sub-árvore que possui o nó escolhido como nó raiz da sub-árvore. Em seu lugar, cria-se uma sub-árvore pelo método *ramped half-and-half*, exatamente como citado no 1º passo do algoritmo de PG, e a mutação está realizada.

(2) Mutação por substituição de regressores: escolhe-se aleatoriamente um regressor de X e realiza-se a substituição deste por outra variável de Ω.

Esta dissertação contemplará as seguintes formas de cruzamento:

(1) Intercâmbio de genes completos dos genitores, de tal forma que o gene de um dos genitores ocupará a posição antiga do gene do outro genitor e vice-versa. São gerados dois descendentes e os genes a serem trocados são escolhidos aleatoriamente.

(2) Cruzamento ao nível intragênico: seleciona-se aleatoriamente um gene de um genitor e outro gene do outro genitor; para cada um dos genes citados, seleciona-se aleatoriamente um nó; a estes nós são associadas sub-árvores, já que cada um dos nós pode ser visto como um nó raiz de suas respectivas sub-árvores; para completar a operação, as sub-árvores dos genitores são trocadas para a geração de dois descendentes.

Nesta dissertação, utilizando-se do *default* do software GPTIPS – plataforma na qual os algoritmos desta dissertação foram construídos, proposta em Searson et al. (2010) –, a taxa de elitismo foi fixada em 5% dos indivíduos por geração.

4
Programação Genética Econométrica

4.1
Introdução e Motivação

Esta dissertação tem como objetivo principal propor modelos de regressão e classificação de elevada acurácia, competitivos frente a algoritmos que desempenhem as mesmas tarefas. Os modelos propostos seguem o arcabouço econométrico do capítulo 2. As estruturas dos modelos de regressão e de classificação são dadas pelas equações (2.1) e (2.18), respectivamente.

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$$

$$P_t = \frac{e^{\boldsymbol{X}_t\boldsymbol{\beta}}}{1 + e^{\boldsymbol{X}_t\boldsymbol{\beta}}} = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$$

O algoritmo gerador de modelos desta dissertação disponibiliza uma família de modelos de regressão ou classificação, em função do conjunto de dados ao qual é aplicado. O modelo de melhor acurácia é aquele superior aos outros de sua família em função de uma métrica de comparação, que deve ser definida em função do tipo de tarefa.

4.1.1
Modelos de Regressão Linear

A estimação por MQO propõe a minimização de uma grandeza associada ao erro.

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = \min_{\boldsymbol{\beta}} \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X\beta})$$

O processo de minimização do SQR é representado por (2.4) e, pelo fato de a função raiz quadrada ser monotonicamente crescente, a grandeza REQM (Raiz do Erro Quadrático Médio), função de SQR, também é minimizada com a minimização de SQR.

$$\mathrm{REQM} = \left(\frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\widehat{\boldsymbol{\beta}})^2 \right)^{1/2} \tag{4.1}$$

A relação entre SQR e REQM é dada por:

$$\mathrm{REQM} = \left(\frac{1}{n} \sum_{t=1}^{n} (y_t - \boldsymbol{X}_t\widehat{\boldsymbol{\beta}})^2 \right)^{1/2} = \left(\frac{1}{n} \mathrm{SQR} \right)^{1/2} \quad \therefore \quad (\mathrm{REQM})^2 = \frac{\mathrm{SQR}}{n} \quad \therefore$$

$$\mathrm{SQR} = n(\mathrm{REQM})^2 \tag{4.2}$$

Embora o REQM seja uma possível medida de ajuste, considerando suas características positivas e negativas – relatadas por Armstrong & Fildes (1995) e Wang & Bovik (2009), entre outros autores –, no domínio de modelos de regressão linear é comum que se utilize o coeficiente de determinação R^2, ou variações dele, como métrica de acurácia.

$$R^2 \equiv 1 - \frac{\mathrm{SQR}}{\mathrm{SQT}} \tag{4.3}$$

A grandeza Somatório dos Quadrados Total (SQT) é dada por $\sum_{t=1}^{n} (y_t - \bar{y})^2$, onde $\bar{y}$ é a média do vetor $\boldsymbol{y}$. O coeficiente R^2 mede o quanto da variabilidade total do conjunto de dados (SQT) consegue ser explicada pelo modelo em questão, em função de sua matriz de regressores $\boldsymbol{X}$, ou seja, representa a fração da variação amostral em que $\boldsymbol{y}$ é explicada por $\boldsymbol{X}$.

Greene (2011) afirma que há problemas com o uso de R^2 como grau de ajuste. Wooldridge (2008) determina que um dos fatores importantes sobre R^2 é o fato de o coeficiente nunca poder diminuir – e, na maior parte dos casos, aumentar – com o acréscimo de variáveis independentes ao modelo de regressão, sejam elas quais forem, i.e., tenham ou não efeito causal sobre a variável dependente. A razão está no comportamento de SQR: do ponto de vista algébrico, SQR nunca aumenta, por definição, quando há adição de variáveis independentes ao modelo. A prova completa deste resultado está em Greene (2011).

Sendo $\boldsymbol{x}_{k+1}$ um vetor de dimensões n x 1, o que se realiza na prática, quando se adiciona uma variável independente genérica x_{k+1} ao conjunto de regressores X do modelo, é a concatenação de $\boldsymbol{x}_{k+1}$ à matriz de regressores $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \dots \boldsymbol{x}_k]$, tornando $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \dots \boldsymbol{x}_k \, \boldsymbol{x}_{k+1}]$ e $X = \{x_1, x_2, \dots, x_k, x_{k+1}\}$.

Levando em consideração a limitação de R^2, é usual que se utilizem algumas de suas variantes – por exemplo, o coeficiente $\overline{R}^2$, chamado de "R^2 ajustado" – como métrica de ajuste, não somente para avaliar um modelo como também para comparar modelos (Greene, 2011): seleciona-se, dentre opções mutuamente excludentes, aquele que apresenta o maior $\overline{R}^2$.

$$\overline{R}^2 = R^2 - \left[(1 - R^2)\frac{k}{n-k-1}\right] \tag{4.4}$$

Segundo Wooldridge (2006), o ponto mais interessante do $\overline{R}^2$ é a penalização à inclusão de variáveis independentes ao modelo caso elas não forneçam melhoria no grau de explicação dos componentes de $\boldsymbol{X}$ a $\boldsymbol{y}$. Diferentemente de R^2, que varia entre 0 e 1, $\overline{R}^2$ pode decrescer quando se adiciona x_{k+1} a X, inclusive podendo o coeficiente assumir valores negativos ou maiores do que 1. De acordo com Greene (2011), $\overline{R}^2$ aumentará, com a adição de x_{k+1}, se a sua contribuição ao ajuste do modelo compensar a perda de uma unidade no número de graus de liberdade $(n - k - 1)$.

A assertiva anterior pode ser formulada pelo seguinte modelo de decisão, segundo Wooldridge (2006): sendo $\hat{\beta}_{k+1}$ estimado por MQO após concatenação de $\boldsymbol{x}_{k+1}$ a $\boldsymbol{X}$ no modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, x_{k+1} aumentará o coeficiente $\overline{R}^2$ se, e somente se, a realização da estatística T, referenciada em (2.39) ao coeficiente

$\hat{\beta}_{k+1}$, for maior do que 1 em valor absoluto. Como já visto, T_{obs} é a realização da variável aleatória T quando os valores de $\hat{\beta}_i$, β_i, $SE(\hat{\beta}_i)$ e $\sqrt{n}$ são substituídos em (2.39). Neste caso ($|T_{obs}| > 1$), x_{k+1} promove um incremento na acurácia ($\overline{\mathrm{R}}^2$) e sua inclusão é justificável. Se $|T_{obs}| < 1$, $\overline{\mathrm{R}}^2$ decresce com a inclusão de x_{k+1}. Se $|T_{obs}| = 1$, $\overline{\mathrm{R}}^2$ não se modifica com a inclusão de x_{k+1}.

O modelo de decisão para acréscimo ou não de x_{k+1} a X apresenta pontos em comum com o modelo de TH descrito na seção 2.3.2. Suas semelhanças se traduzem na presença de uma estatística de teste com distribuição conhecida e em critérios de decisão representados pelas regiões de rejeição e não-rejeição. Enquanto o TH de 2.3.2 determina $\boldsymbol{H}_0$ e $\boldsymbol{H}_1$ explicitamente, a decisão de acréscimo de x_{k+1} a X não explicita hipóteses sobre quaisquer coeficientes populacionais, sendo esta uma diferença entre os processos.

Para efeito de exemplificação, com n suficientemente grande, $n \gg k$, $\alpha = 5\%$ e $T_{obs} = 0{,}40$, x_{k+1} não seria adicionado a X, pois $|T_{obs}| = 0{,}40 < 1{,}00$, e também não seria considerado estatisticamente significante, pois $|T_{obs}| < 1{,}96 = t_{\alpha/2}(n-k-1)$. Se $T_{obs} = 1{,}60$, x_{k+1} seria adicionado ao modelo, mas não seria estatisticamente significante. Se $T_{obs} > 1{,}96$, há uma situação favorável ao modelo, pois há acréscimo de acurácia com significância estatística.

Diz-se que x_i é estatisticamente significante se há rejeição de $\boldsymbol{H}_0$ para os modelos de TH de (2.29). A significância estatística é fundamental para que se possa atribuir relações de causa e efeito entre variáveis, evitando que se tomem efeitos puramente aleatórios como causas de eventos de interesse. Há diversas possíveis razões pelas quais x_i não seja estatisticamente significante em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, ao nível α: uma delas é que é possível que x_i seja estatisticamente significante a um nível de significância maior promovendo, entretanto, um TH com nível de confiança menor; como o TH é realizado sobre β_i, um parâmetro populacional, é possível que x_i não pertença de fato ao conjunto de melhores regressores que infiram o comportamento de y.

$\overline{\mathrm{R}}^2$ pode ser escrito como em (4.5), supondo $c = k/(n-k-1)$:

$$\overline{\mathrm{R}}^2 = \mathrm{R}^2(1+c) - c \tag{4.5}$$

Há uma relação entre REQM e $\overline{\mathrm{R}}^2$:

$$\overline{\mathrm{R}}^2 = \mathrm{R}^2(1+c) - c = \left(1 - \frac{n(\mathrm{REQM})^2}{\mathrm{SQT}}\right)(1+c) - c$$

$$\overline{\mathrm{R}}^2 = 1 - \left(\frac{n(\mathrm{REQM})^2}{\mathrm{SQT}}\right)(1+c) \tag{4.6}$$

A partir de (4.6), conclui-se que a maximização de $\overline{\mathrm{R}}^2$ é implicada pela minimização de REQM que, por sua vez, é minimizada quando se minimiza SQR.

Dentre REQM, R^2 e $\overline{\mathrm{R}}^2$, seria o $\overline{\mathrm{R}}^2$ a métrica utilizada para, inicialmente, definir o conceito de acurácia para a tarefa de regressão e, em seguida, comparar os modelos de regressão linear. O coeficiente $\overline{\mathrm{R}}^2$ seria escolhido pelo fato de sua maximização estar diretamente relacionada ao processo de estimação de parâmetros (MQO), além de sua natureza parcimoniosa permitir que somente haja aumento em seu valor se o acréscimo de regressores for justificável do ponto de vista estatístico. Embora seja questionado por não penalizar de uma maneira mais rigorosa os graus de liberdade – Amemiya (1985) sugere uma série de outras grandezas que o fazem –, o $\overline{\mathrm{R}}^2$ costuma ser sugerido como uma primeira alternativa ao R^2 e goza de boa aceitação como métrica de ajuste e comparação de modelos nos pacotes econométricos (Davidson & MacKinnon, 2003).

Entretanto, se $k > n$, então $c < 0$. Com isso, $\overline{\mathrm{R}}^2$ pode assumir valores muito altos para modelos que apresentam R^2 muito baixo. Como exemplo, supondo-se um modelo com as especificações $k = 10$, $n = 8$ e $\mathrm{R}^2 = 0{,}10$, $\overline{\mathrm{R}}^2$ assumiria o valor $3{,}10$. Claramente, o valor de $\overline{\mathrm{R}}^2$ não condiz com a capacidade do modelo em explicar y – seu valor é alto somente pelo fato de k ser maior do que n. Embora, a princípio, suponha-se que estas conjecturas não ocorram em situações reais, elas de fato ocorrem: quando se utiliza o processo de validação cruzada para validação de modelos e/ou o conjunto de dados é relativamente pequeno (n é pequeno), é possível que modelos com boa acurácia no conjunto de treino apresentem desempenho ruim no conjunto de validação ou até mesmo no conjunto de treino (caso $n < 50$). Portanto, $\overline{\mathrm{R}}^2$ não poderia ser utilizado nestas situações, limitando a utilização do algoritmo de geração de modelos.

Construir modelos de regressão de acurácia elevada é o objetivo desta dissertação. Para que seja cumprido, será proposto um algoritmo gerador de modelos de regressão linear que se utilize do REQM (já minimizado pela estimação por MQO, que minimiza o SQR, que ocasiona maximização de $\overline{R}^2$) como métrica de comparação entre modelos. Embora $\overline{R}^2$ não seja explicitamente utilizado como métrica de comparação entre os modelos ao longo da evolução, devido à razão já mencionada, ele pode ser utilizado ao término da evolução como métrica de comparação com o *benchmark* proposto.

Para que se cumpra o objetivo, será ferramenta fundamental a adição de regressores estatisticamente significantes à X de cada um dos modelos propostos, ao nível de significância de 5%, aos moldes de (2.29) para TH. Ao considerar o limiar $t_{\alpha/2}(n-k-1)$, em substituição ao valor unitário, tanto para realizar TH em β_{k+1} quanto para decidir se é correto ou não acrescentar x_{k+1} (do ponto de vista da maximização de $\overline{R}^2$), preza-se por modelos que possuam regressores altamente colaborativos com o grau de explicação de y, além de serem estaticamente significantes. Tal decisão revela uma característica conservadora do método, ao somente permitir em X os regressores que sejam de fato agregadores à acurácia do modelo, aumentando o limiar de decisão para adição de x_{k+1} do valor $1{,}00$ para o valor $1{,}96$.

O algoritmo de geração de modelos de regressão linear se utiliza da prova matemática relacionada ao acréscimo de x_{k+1} a X (o acréscimo de regressores nunca diminui o R^2, consequentemente nunca aumenta o REQM) e da condição necessária para que x_{k+1} seja estatisticamente significante e, por consequência, gerador de aumento a $\overline{R}^2$. Ou seja, à X deve ser acrescentado o maior número de regressores que sejam estatisticamente significantes.

Idealmente, o processo gerador de regressores deve ser independente de domínio – caso contrário, considerando uma série de dez conjuntos de dados, supondo que sejam de naturezas distintas, seriam necessários um ou mais especialistas por conjunto de dados para propor tais regressores.

A PG é a ferramenta mais indicada para desempenhar tal tarefa, porque não somente cumpre com a geração de possíveis regressores a X, permitindo aos modelos aproveitamento considerável do espaço de busca, como também realiza

todo o processo de evolução de modelos, utilizando-se de seus operadores genéticos para propor novas regressões.

4.1.2
Modelos de Regressão Não Linear: Classificação Binária

A seção 4.1.1 caracterizou em sua totalidade o conjunto de razões que sustentam o mecanismo gerador de modelos de regressão.

A equação (2.22) mostra a função $l(\boldsymbol{\beta})$, a ser maximizada no processo de estimação de $\boldsymbol{\beta}$ por MV.

$$l(\boldsymbol{\beta}) = \sum_{t=1}^{n} [y_t \log\big(\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})\big) + (1 - y_t) \log(1 - \Lambda(\boldsymbol{X}_t\boldsymbol{\beta}))]$$

$\widehat{\boldsymbol{\beta}}$, estimador de MV, não maximiza uma medida de ajuste (tal como $\overline{\mathrm{R}}^2$) ou minimiza determinado tipo de erro (tal como REQM) – $\widehat{\boldsymbol{\beta}}$ maximiza $l(\boldsymbol{\beta})$. Embora pareça haver ganhos em métricas de acurácia associadas a algum tipo de erro quando $l(\boldsymbol{\beta})$ é maximizada, segundo Greene (2011), de maneira genérica, permanece como uma questão interessante aos pesquisadores privilegiar a minimização do erro (tal como o MQO) ou a obtenção de bons estimadores (tal como a MV) como decisão frente à definição de seus modelos.

Particularmente em relação à classificação binária por RL, Davidson & MacKinnon (2003) afirmam que os modelos logit classificarão perfeitamente um conjunto de dados para um dado $\boldsymbol{\beta}^*$ se $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}^*) = 1$ quando $y_t = 1$ e $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta}^*) = 0$ quando $y_t = 0$. A condição anterior ocorre somente se $\boldsymbol{X}_t\boldsymbol{\beta}^* = \infty$, quando $y_t = 1$, e $\boldsymbol{X}_t\boldsymbol{\beta}^* = -\infty$, quando $y_t = 0$. Satisfeitas as condições anteriores, $l(\boldsymbol{y}, \boldsymbol{\beta}^*) = 0$ (ou próximo a zero) e $\boldsymbol{\beta}^*$ é o estimador de MV tal que $\boldsymbol{X}_t\boldsymbol{\beta}^*$ é nomeado classificador perfeito, promovendo uma separação perfeita para y supondo X.

Portanto, é possível inferir que a maximização de $l(\boldsymbol{\beta})$ proporcionará minimização do percentual de classificações incorretas – que é uma métrica associada a erro – quando X é selecionado da melhor maneira possível. Analogamente, maximiza-se o percentual de classificações corretas quando a solução de (2.22) é $\boldsymbol{\beta}^*$.

O mecanismo gerador de modelos para a tarefa de regressão seria idealmente estruturado tendo o $\overline{R}^2$ como medida de acurácia e comparação de modelos. Devido a possíveis problemas oriundos de sua forma de cálculo, foi o REQM a grandeza escolhida em sua substituição. O algoritmo gerador de modelos para a tarefa de classificação será exposto a algo semelhante: não será a função $l(\boldsymbol{\beta})$ a métrica de acurácia a ser minimizada e usada para comparar modelos – será utilizado o percentual de classificações incorretas frente ao total de classificações realizadas (grandeza que será denominada por "$\%\_inc$"). Há uma única razão que justifica tal fato: todos os *benchmarks* utilizados para comparar os modelos de classificação gerados nesta dissertação, em distintos conjuntos de dados, têm como métrica de comparação $\%\_inc$. Como avaliado por Davidson & MacKinnon (2003), a minimização de $\%\_inc$ possui uma relação direta com a maximização de $l(\boldsymbol{\beta})$.

Continua sendo ferramenta fundamental para que se cumpra o objetivo da dissertação a adição de regressores estatisticamente significantes à X de cada um dos modelos propostos, ao nível de significância de 5%, nos moldes de (2.29) para TH. Observa-se que os TH serão aplicados considerando-se a natureza da tarefa: se é de regressão, a teoria na seção 2.3.2 para TH deve ser aplicada; caso contrário, a teoria presente na seção 2.3.3.

Não há prova matemática que comprove que mais regressores em X fazem com que o método de MV apresente $l(\boldsymbol{\beta})$ maior e, consequentemente, menor $\%\_inc$. Prova-se, sob as Hipóteses 1 a 5 (seção 2.3.2) para a estimação por MQO e sob normalidade de y para a estimação por MV, que os estimadores de MQO e MV coincidem (Koopmans, 1950). Entretanto, y não tem distribuição normal e o resultado não se aplica.

Para justificar a proposta de inclusão de regressores estatisticamente significantes a X nos modelos de classificação, para efeito de aumento em $l(\boldsymbol{\beta})$, intuitivamente analisa-se que o aumento do conjunto X tende a caracterizar melhor o grupo de variáveis independentes que explicam o comportamento de $\Pr(y_t = 1|\boldsymbol{X})$.

Tal análise será suficiente para que o critério de inclusão de regressores a X permaneça, para a geração de modelos de classificação, semelhante ao que

possui o mecanismo gerador de modelos de regressão linear, embora não haja prova matemática que sustente esta decisão.

4.2
Hipóteses

A teoria proposta no capítulo 2, relativa à estimação e TH em modelos de regressão linear e não linear, utilizada nesta dissertação pelo algoritmo de geração de modelos aplicados a tarefas de regressão e classificação, só pode ser plenamente utilizada se as hipóteses que sustentam a teoria citada se mantiverem para os conjuntos de dados tratados.

A estimação por MQO para modelos de regressão linear é uma solução geométrica, requisitando apenas que $\boldsymbol{X}$ tenha posto cheio: a decomposição QR realiza essa checagem e inabilita a presença de regressores linearmente dependentes na estimação de $\widehat{\boldsymbol{\beta}}$.

A estimação por MV para modelos logit pressupõe que $\boldsymbol{y}$ seja um vetor de realizações de variáveis aleatórias dicotômicas com distribuição de Bernoulli e que cada uma das n realizações seja oriunda de uma variável aleatória y_t com a mesma distribuição. A priori, qualquer AA de variáveis aleatórias dicotômicas y_t pode ser modelada como uma AA de variáveis aleatórias com distribuição de Bernoulli, pois a natureza de y_t permite que assim seja feito, dada a definição da variável aleatória de Bernoulli (Casella & Berger, 2011). Supondo a assertiva anterior, pode-se estimar $\widehat{\boldsymbol{\beta}}$ tal que $\%\_inc$ não seja satisfatório: isto não infere que a caracterização da distribuição de probabilidades de y_t tenha sido incorreta. É possível que X não comporte regressores que tendam a explicar melhor o comportamento de y, que o MN (Método de Newton) tenha fornecido um ótimo local, ou ambas as razões. Independentemente de qual seja a razão, a especificação de y_t como variável aleatória de Bernoulli é justificável, permitindo a utilização da MV como método de estimação de $\widehat{\boldsymbol{\beta}}$.

TH para modelos logit são válidos como proposto no capítulo 2 porque se considera que n é suficientemente grande de tal forma que a análise assintótica, que propõe distribuição assintoticamente normal para $\widehat{\boldsymbol{\beta}}$, seja válida e permita TH

com a distribuição da estatística de teste correspondente. Não é comum a modelagem de heterocedasticidade para modelos logit em trabalhos empíricos.

TH para modelos de regressão linear supõem que as Hipóteses 1 a 5, citadas na seção 2.3.2, se sustentem. As Hipóteses 1, 2 e 3 se sustentam por linearidade dos parâmetros em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, amostragem aleatória e uso da decomposição QR para estimar $\widehat{\boldsymbol{\beta}}$ por MQO. O estimador de $\widehat{\boldsymbol{\beta}}$ por MQO tem distribuição normal assintótica, como provado em Eicker (1963), não havendo necessidade de normalidade para $\boldsymbol{u}|\boldsymbol{X}$ (Hipótese 5). Embora a teoria proposta para TH em modelos de regressão linear no capítulo 2 suponha amostras finitas, os resultados para TH em $\boldsymbol{\beta}$ (supondo normalidade de $\boldsymbol{u}|\boldsymbol{X}$) serão semelhantes aos resultados para análises assintóticas quando n é suficientemente grande, porque, como já visto, $T_{n_{gl}} \xrightarrow{d} N(0,1)$.

A ausência de correlação serial é atendida com amostragem aleatória, satisfazendo a um dos requisitos da Hipótese 4. A homocedasticidade, o outro requisito que deve ser atendido na Hipótese 4, será abordada na seção seguinte.

4.2.1
Homocedasticidade

No capítulo 2, mostrou-se que $Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) = \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$, a partir da hipótese $Var\,(\boldsymbol{u}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}_n$. Pode-se reescrever $Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X})$ como:

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\right] \\ &= E[((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})((\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u})^{\mathrm{T}}] \\ &= E[(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}] \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}E[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}|\boldsymbol{X}]\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}, \end{aligned} \tag{4.7}$$

onde:

$$E[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}|\boldsymbol{X}] = \begin{bmatrix} E[u_1^2|\boldsymbol{X}] & \cdots & E[u_1u_n|\boldsymbol{X}] \\ \vdots & \ddots & \vdots \\ E[u_nu_1|\boldsymbol{X}] & \cdots & E[u_n^2|\boldsymbol{X}] \end{bmatrix}. \tag{4.8}$$

A ausência de correlação serial é atendida com amostragem aleatória, logo, qualquer elemento fora da diagonal principal de $E[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}|\boldsymbol{X}]$ é nulo.

$$E[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}|\boldsymbol{X}] = \begin{bmatrix} E[u_1^2|\boldsymbol{X}] & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & E[u_n^2|\boldsymbol{X}] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} \tag{4.9}$$

A hipótese de homocedasticidade, que propõe $\sigma_i^2 = \sigma^2$, nem sempre se verifica em situações práticas. Quando $\sigma_i^2 \neq \sigma^2$, a situação é de heterocedasticidade, que não interfere na estimação de $\widehat{\boldsymbol{\beta}}$ por MQO, mas interfere em TH para $\boldsymbol{\beta}$ porque $Var\,(\boldsymbol{u}|\boldsymbol{X})$ pode ser maior ou menor sob heterocedasticidade.

White (1980) prova que $\boldsymbol{X^T\widehat{u}\widehat{u}^T X}$ é um estimador consistente, mas viesado, de $\boldsymbol{X}^{\mathrm{T}}E[\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}|\boldsymbol{X}]\boldsymbol{X}$. O termo $\boldsymbol{\widehat{u}\widehat{u}^T}$ é denominado variância robusta à heterocedasticidade, representando um estimador válido na presença de homocedasticidade ou heterocedasticidade desconhecida. Ou seja, pode-se realizar TH independentemente do tipo de variância (homocedasticidade ou heterocedasticidade) ou da forma de variância (refere-se à heterocedasticidade, que pode apresentar uma estrutura) presente na população.

A partir dos trabalhos de White (1980), Eicker (1967) e Huber (1967), que afirmam ser possível obter estimadores deste gênero, constituiu-se uma das maneiras de tratar heterocedasticidade em modelos de regressão linear, permitindo-se realizar TH para $\boldsymbol{\beta}$: substitua $\sigma^2\boldsymbol{I}_n$ por $\boldsymbol{\widehat{u}\widehat{u}^T}$ em $Var\left(\widehat{\boldsymbol{\beta}}\middle|\boldsymbol{X}\right)$. O termo $\sigma^2\boldsymbol{I}_n$ é nomeado variância homocedástica e $\boldsymbol{\widehat{u}\widehat{u}^T}$ é nomeada variância heterocedástica ou variância de White.

Para 5.000 modelos de regressão linear obtidos para cada conjunto de dados relacionado à tarefa de regressão desta dissertação, realizaram-se dois TH, seguindo (2.29), para cada um dos regressores desses modelos: um com a variância homocedástica e outro com a variância de White. Os resultados são apresentados na Tabela 4.1.

Tabela 4.1 – Resultados para TH em Regressores com Variâncias distintas

Conjunto de Dados	TH: Variâncias		
	média	mediana	#reg
1	92,31%	94,74%	105.515
2	89,87%	91,91%	101.222
3	90,08%	94,75%	72.526
4	88,10%	93,00%	112.842
5	91,20%	91,82%	61.463

Na tabela 4.1, “média” representa o percentual médio de similaridade entre os TH realizados para todos os regressores presentes em 5.000 modelos. Por exemplo, para o conjunto de dados 1, houve similaridade média em 92,31% dos TH realizados em 105.515 regressores presentes em 5.000 indivíduos – ou seja, o TH com variância homocedástica e o TH com variância de White apresentaram o mesmo resultado em 92,31% dos casos. “Mediana” representa a mediana do conjunto de TH. “#reg” é o total de regressores avaliados nos testes de hipóteses individuais.

A tabela evidencia, empiricamente, a partir da observação da média e mediana, que a utilização de uma ou outra variância modifica somente marginalmente o resultado de TH para $\boldsymbol{\beta}$, tendo como base os conjuntos de dados desta dissertação e o experimento proposto com 5.000 indivíduos. Adota-se, portanto, a variância homocedástica como base para TH em $\boldsymbol{\beta}$ nesta dissertação.

Após a introdução dos principais elementos que motivaram a construção do mecanismo gerador de modelos de regressão e classificação proposto nesta dissertação, nomeado Programação Genética Econométrica (PGE), além da análise de sustentação das hipóteses que estruturaram o arcabouço teórico, será apresentada a PGE como mecanismo que realiza o processo de evolução de modelos econométricos de regressão e classificação.

4.3
O Modelo de PGE

A PGE é o algoritmo de PG, que evolui modelos econométricos, proposto nesta dissertação. Portanto, esta seção se utilizará do mesmo roteiro proposto no capítulo 3, para a PG, como forma de apresentação da PGE.

A Figura 4.1 mostra o pseudocódigo do algoritmo de PGE, quando se substitui o termo “programas” por “modelos” ou “regressões”, na Figura 3.1. Se a variável dependente y é real, a PGE realizará a evolução de modelos da forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. Caso y seja binária, os modelos serão da forma $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$, como em (2.18).

```
1: Crie aleatoriamente uma população inicial de modelos,
   a partir das primitivas disponíveis.
2: REPITA
3:     Determine a acurácia de cada modelo.
4:     Selecione um ou dois modelos da população, em função
       de probabilidades determinadas pela acurácia, para que
       façam parte do processo de criação da nova população,
       utilizando os operadores de mutação e/ou cruzamento.
5:     Crie novos modelos, aplicando os operadores citados,
       de acordo com probabilidades do experimento (são distintas
       das probabilidades determinadas pela acurácia).
6: ATÉ QUE: uma solução aceitável seja encontrada ou outra
   condição de parada seja atingida (por exemplo, o número
   máximo de gerações).
7: RETORNE: o modelo de melhor acurácia.
```

Figura 4.1 – Pseudocódigo do algoritmo de PGE

Este pseudocódigo servirá de base para a descrição da PGE.

4.3.1
Representação

Na PGE, os programas (regressões, modelos ou indivíduos) são representados como na Figura 3.3.

Qualquer constante em um indivíduo da PGE é proveniente da estimativa de $\widehat{\boldsymbol{\beta}}$, oriundo da estimação por MQO, se a tarefa for de regressão ou da

maximização de $l(\boldsymbol{\beta})$, se a tarefa for de classificação. Portanto, Ω nesta dissertação é composto somente por variáveis.

4.3.2
1º Passo: Criação da População Inicial

A PGE se utilizará de uma versão probabilística do método *ramped half-and-half* para geração da população inicial. Internamente, a PGE criará a população da seguinte forma: supondo que a altura máxima permitida da árvore seja 5 e o tamanho da população seja igual a 100, 20 indivíduos serão gerados a cada altura de árvore, de 1 à 5 (altura máxima permitida). Do total da população, metade dos indivíduos será gerada pelo método *full* e a outra metade pelo método *grow*. Cópias múltiplas de genes em um indivíduo da população inicial são proibidas, embora esta restrição não ocorra ao longo da evolução.

Um conjunto de dados desta dissertação origina o seu conjunto de terminais, Ω, que dispõe de K variáveis de entrada: $x_1, x_2, \dots, x_I, \dots, x_K$. Cada conjunto de dados originará o seu próprio Ω. Suponha, somente para o exemplo em seguida, que haja um conjunto de dados com Ω contendo trinta variáveis de entrada $(K = 30)$. Estas variáveis serão os possíveis regressores de cada indivíduo da população inicial. Por exemplo, é possível que o método *ramped half-and-half* gere o indivíduo da Figura 4.2. O método atua em cada gene, individualmente, possibilitando a criação de genes com alturas distintas.

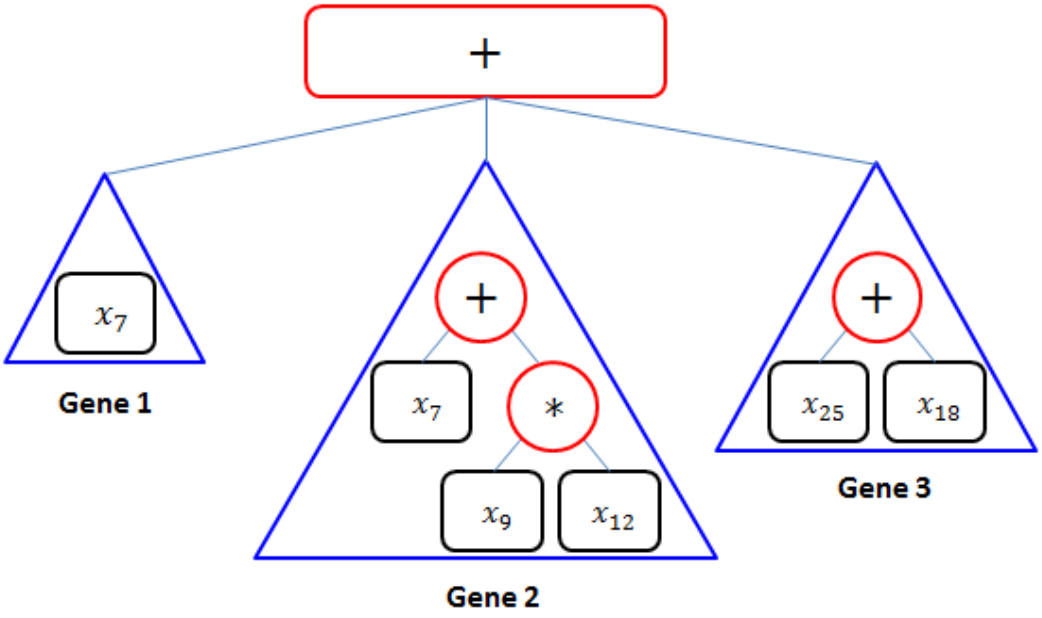

Figura 4.2 – Indivíduo multigênico típico de um experimento de PGE

Fonte: adaptado de Gandomi & Alavi (2011)

É fundamental atentar que as variáveis $x_I, I \in [1, K]$, e $x_i, i \in [1, k]$ representam grupos distintos. Enquanto o 1º retrata o conjunto de variáveis de entrada que compõem Ω, o 2º faz referência à matriz de regressores $\boldsymbol{X} \equiv [\boldsymbol{x}_1 \ldots \boldsymbol{x}_k]$, com cada coluna $\boldsymbol{x}_i$ de $\boldsymbol{X}$ sendo um vetor de dimensões n x 1, de um modelo como em $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ ou $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$.

Quando o método *hamped half-and-half*, na construção de um indivíduo, aleatoriamente seleciona uma variável x_I de Ω para ser um regressor x_i que compõe X para $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ ou $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$, supõe-se que há uma parametrização de I à i, fazendo com que a teoria econométrica seja aplicada utilizando-se da notação do capítulo 2.

Posteriormente, será mostrado que o indivíduo exemplificado pela Figura 4.2, assim como qualquer outro indivíduo de experimentos de PGE, será estruturado como um modelo da forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, caso a tarefa associada ao conjunto de dados seja de regressão, ou $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$, caso contrário.

4.3.3
2º Passo: Estrutura de Repetição

Nos experimentos de PGE, a condição de parada do algoritmo para todos os conjuntos de dados é o número máximo de gerações proposto.

4.3.4
3º Passo: Determinação e Cálculo da Acurácia

Ω é composto somente por variáveis de seu conjunto de dados. Como visto anteriormente, qualquer indivíduo de um experimento de PGE será estruturado como um modelo da forma $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, caso a tarefa associada ao conjunto de dados seja de regressão, ou $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$, caso contrário. Os formatos de modelos citados somente necessitam das operações de soma e multiplicação para que possam ser construídos. Logo, o conjunto de funções da PGE é composto somente pelas funções de soma e multiplicação.

As funções do conjunto de funções da PGE possuem a propriedade de consistência de tipo: tanto os argumentos quanto a saída das funções são números reais. A PGE não apresenta possibilidade de erros em tempo de execução, pelo fato das operações de soma e multiplicação serem definidas para todos os reais. Logo, a PGE atende à propriedade de segurança na avaliação. A PGE cumpre com a propriedade de suficiência ao propor os modelos $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ e $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ para as tarefas de regressão e classificação, respectivamente. Portanto, pode-se dizer que a PGE é um algoritmo de PG apto a realizar a evolução de indivíduos efetivamente.

Para a PGE, o espaço de busca é o número total de modelos, n_{mod}, que podem ser gerados em um experimento. Por sua vez, n_{mod} é função do número de regressores que podem ser criados, n_{reg}.

Usa-se o termo “criação” de regressores pelo fato de os regressores em um indivíduo não serem somente (e no máximo) as K variáveis de entrada disponíveis em Ω. Considera-se um regressor para X qualquer combinação (necessariamente via multiplicação) de variáveis de Ω. Por exemplo, a multiplicação das variáveis

x_1 e x_2 de Ω, x_1x_2, é interpretada como um regressor em X. O fato de x_1x_2 pertencer à X não impossibilita que x_1 e/ou x_2 também pertençam.

Apresentam-se abaixo as expressões para os cálculos de n_{reg} e n_{mod}.

$$n_{reg} = \sum_{q_{var}=1}^{K} \frac{(K-1+q_{var})!}{(K-1)!\,q_{var}!} \tag{4.10}$$

$$n_{mod} = \sum_{q_{reg}=1}^{n_{reg}} \frac{n_{reg}!}{(n_{reg}-q_{reg})!} \tag{4.11}$$

Em (4.10), q_{var} é a quantidade de variáveis de Ω necessárias para se criar um regressor, e q_{reg} é a quantidade de regressores utilizada para se criar um modelo. A expressão (4.10) é o somatório das possíveis combinações com repetições de K variáveis pertencentes a Ω, q_{var} a q_{var}. A expressão (4.11) é o somatório dos possíveis arranjos de n_{reg} regressores, q_{reg} à q_{reg}. Supondo $K = 3$ para um conjunto de dados, n_{mod} é da ordem de 10^{17}.

Na PGE, a função objetivo para tarefas de regressão será a REQM; para tarefas de classificação, será $\%\_inc$.

O cálculo da acurácia não é realizado utilizando-se diretamente a estrutura multigênica da Figura 3.3. O processo de "modificação" de indivíduos para cálculo da acurácia será mostrado pela sequência das Figuras 4.3 à 4.5, considerando-se novamente o indivíduo da Figura 4.2. Os indivíduos não são de fato modificados: as transformações propostas são realizadas somente para que se calcule a acurácia. A estrutura multigênica original do indivíduo, como na Figura 3.3, permanece inalterada e será ela a designada às etapas seguintes do algoritmo, particularmente às fases de seleção e criação de nova população (por mutação, cruzamento e elitismo).

Inicialmente, deve-se escrever explicitamente cada gene como um somatório de variáveis de Ω e/ou combinações de variáveis de Ω, como na Figura 4.3.

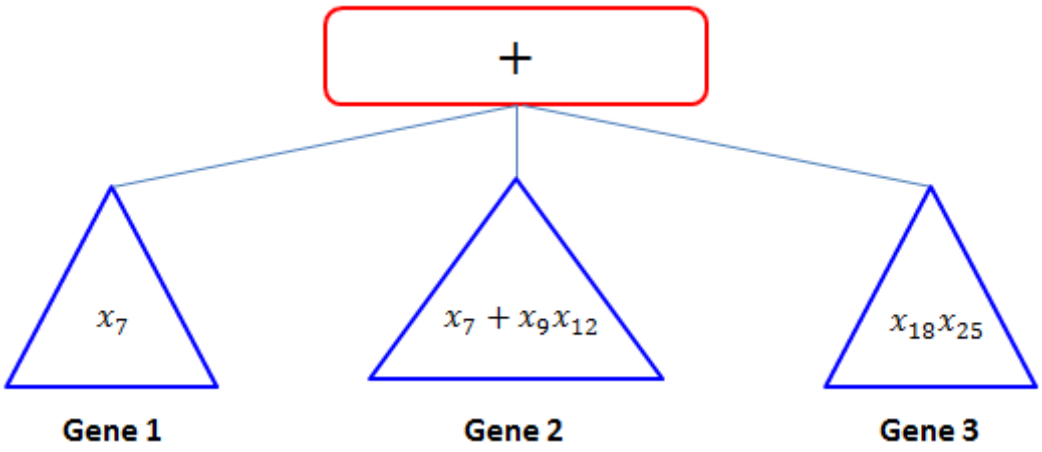

Figura 4.3 – Cálculo da acurácia: 1ª etapa

Em seguida, os regressores serão retirados de seus respectivos genes. Não há qualquer interesse em estimativas de $\widehat{\boldsymbol{\beta}}$ para genes, e sim para regressores, devido ao potencial problema de multicolinearidade perfeita, que pode ocorrer caso seja realizada a estimação no domínio dos genes. Supondo um indivíduo formado por dois genes, um deles contendo somente o regressor x_1 e o outro contendo exclusivamente o regressor $-x_1$, haverá multicolinearidade perfeita, potencialmente interferindo nos resultados para TH sobre $\boldsymbol{\beta}$, entre outras consequências, como relata Wooldridge (2008).

Sob o domínio da PGE, a multicolinearidade não apresenta uma preocupação quanto ao desempenho do algoritmo gerador de modelos. A estimação de $\widehat{\boldsymbol{\beta}}$ por MQO em modelos de regressão linear, realizada pela decomposição QR, utiliza-se do fato de que, quando o posto de $\boldsymbol{X}$ não é cheio e há m colunas de $\boldsymbol{X}$ que sejam linearmente dependentes das $k - m$ colunas restantes, o algoritmo é modificado de tal forma que $\boldsymbol{Q}$ tenha $k - m$ colunas e $\boldsymbol{R}$ tenha dimensões $(k - m)$ x k. Dessa forma, a estimativa de $\widehat{\boldsymbol{\beta}}$ torna-se solução única do algoritmo quando se arbitra que os coeficientes dos m regressores linearmente dependentes sejam iguais à zero. A estimação de $\widehat{\boldsymbol{\beta}}$ por MV em modelos logit é realizada aplicando-se o MN às equações que retratam as condições de 1ª ordem da maximização de $l(\boldsymbol{\beta})$. Tal processo resolve iterativamente sistemas de equações em $\boldsymbol{\beta}$ – caso o posto de $\boldsymbol{X}$ não seja cheio, a otimização sequer é realizada.

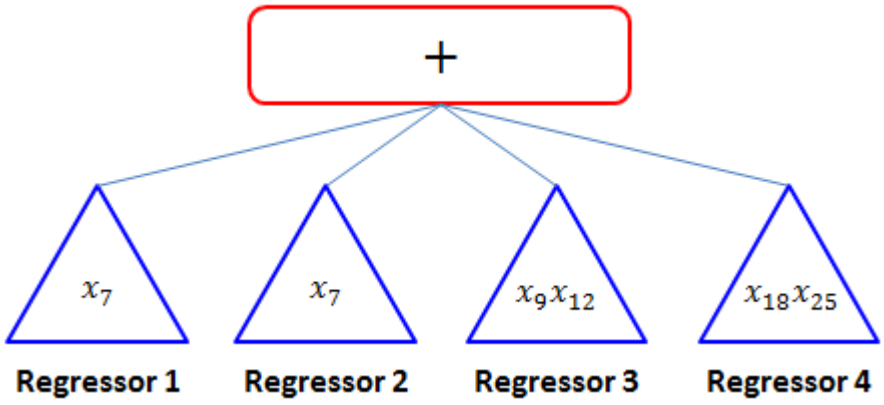

Figura 4.4 – Cálculo da acurácia: 2ª etapa

Prosseguindo com o cálculo da acurácia do indivíduo exemplificado, observa-se que, pelo fato dos regressores 1 e 2 representarem o mesmo regressor (Figura 4.4), somente um deles será necessário. Caso todos os regressores fossem distintos entre si, não haveria qualquer exclusão.

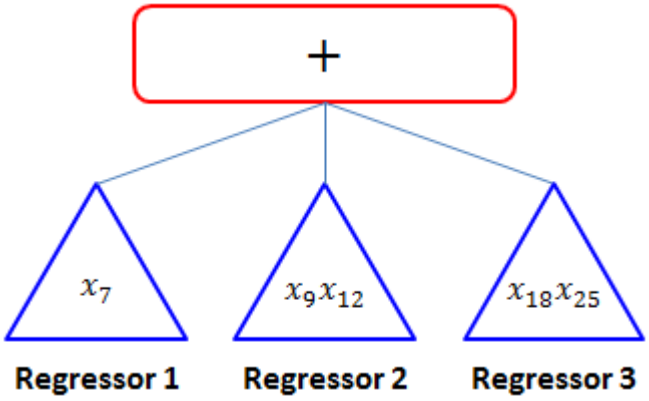

Figura 4.5 – Cálculo da acurácia: 3ª etapa

A partir da Figura 4.5, é possível escrever o indivíduo como um modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ ou $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$, pois $\boldsymbol{X\beta} = \beta_1 x_7 + \beta_2 x_9 x_{12} + \beta_3 x_{18} x_{25}$. Se o indivíduo fictício for proposto pela PGE em uma tarefa de regressão, utiliza-se o modelo $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ para estimação de $\widehat{\boldsymbol{\beta}}$ por MQO. Depois de estimado $\widehat{\boldsymbol{\beta}}$, avalia-se quais regressores em X são estatisticamente significantes. Os que não o são, de acordo com o TH proposto no capítulo 2, serão retirados de X. Realiza-se

uma nova estimação somente com os regressores estatisticamente significantes em X, chegando à $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$. O cálculo de REQM é realizado em função de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$.

É importante atentar novamente que a estrutura multigênica original do indivíduo permanece inalterada e será ela a designada às etapas seguintes do algoritmo. O ponto favorável de a PGE fazer estimação ao nível dos regressores e manter a estrutura multigênica para o prosseguimento da evolução é o fato dos genes de um indivíduo poderem se recombinar, a partir do cruzamento, ou mutar a fim de produzir outros regressores com significância estatística. Eliminar permanentemente genes e/ou regressores dos indivíduos ocasionaria perda de diversidade genética na população. Como exemplo, supondo que x_1 não seja estatisticamente significante em uma dada regressão, tal fato não impede que x_1^2 o seja em outra regressão. Se x_1 ou o gene que comporta x_1 é eliminado, torna-se menor a chance de se obter x_1^2 em algum outro indivíduo, por mutação ou cruzamento.

Se o indivíduo for proposto pela PGE em uma tarefa de classificação, utiliza-se o modelo $\Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ para estimação de $\widehat{\boldsymbol{\beta}}$ por MV. Da mesma forma como na descrição para a tarefa de regressão, será $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$ o vetor de estimativas utilizado para cálculo da acurácia. $\Lambda\left(\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}\right)$, o vetor de probabilidades estimadas através de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}$, é utilizado para determinar $\hat{y}_t$ através da seguinte regra: se $\boldsymbol{X}_t\widehat{\boldsymbol{\beta}}_{\boldsymbol{2}} \geq 0{,}5$, $\hat{y}_t = 1$; caso contrário, $\hat{y}_t = 0$. O limiar de probabilidade de valor 0,5 foi escolhido para que não haja viés à classificação de uma ou outra classe. O limiar também independe do conjunto de dados ser desbalanceado ou não. Com $\hat{y}_t$, calcula-se a acurácia do modelo, dada por $\%\_inc$.

4.3.5
4º Passo: Seleção

A PGE utiliza o método de torneio para seleção de indivíduos, com $n_{torneio} = 7$. São aleatoriamente escolhidos da população $n_{torneio}$ indivíduos, permitindo que se selecione o mesmo indivíduo mais de uma vez.

Na PGE, utiliza-se a seleção por torneio com uma variante de pressão lexicográfica: inicialmente, todos os indivíduos da população são separados em grupos de acordo com a sua acurácia, considerando-se a partir dessa divisão que

indivíduos do mesmo grupo possuam acurácias iguais; em seguida, realiza-se a seleção por torneio; como haverá indivíduos com acurácias iguais, dado que essa situação foi induzida, será considerado o vencedor aquele com a maior quantidade de regressores estatisticamente significantes.

Enquanto a pressão lexicográfica proposta por Luke & Panait (2002) atua parcimoniosamente sobre a evolução controlando o número de nós, a variante proposta para a PGE também tem natureza parcimoniosa, ao atuar sobre o número de regressores estatisticamente significantes. A razão pela qual se desejam regressores nestas características já foi analisada.

A diferença entre a pressão lexicográfica proposta por Luke & Panait (2002) e a pressão lexicográfica utilizada pela PGE está na variável que cada uma controla e no momento em que a pressão lexicográfica é aplicada: em Luke & Panait (2002), somente os indivíduos de mesma acurácia, após terem sido selecionados para o torneio, competem entre si em relação ao número de nós – vencerá aquele que tiver o menor número de nós. Na pressão lexicográfica utilizada pela PGE, antes mesmo da seleção os indivíduos já são categorizados em função da sua acurácia para competirem no torneio, posteriormente. A fim de não linearizar a população, propôs-se que houvesse somente dois indivíduos por classe para categorização de indivíduos em função da acurácia.

4.3.6
5º Passo: Mutação, Cruzamento e Elitismo

A PGE se utilizará das formas de mutação e cruzamento descritas para a PG, assim como do percentual para taxa de elitismo, fixado em 5%.

Na PGE, as probabilidades de ocorrência dos operadores de mutação e cruzamento são variantes ao longo do experimento. Davis (1989) afirma que, embora grande parte das implementações de algoritmos genéticos (veja Holland, 1992) mantenham fixas as probabilidades de ocorrência de seus operadores genéticos (probabilidades do experimento) durante o experimento, essas probabilidades deveriam ser variantes, baseando-se principalmente na capacidade dos operadores em fornecer filhos com acurácia melhor do que seus pais.

Silva (2007) propõe, no software livre GPLAB, a utilização de um processo de adaptação automática para as probabilidades de ocorrência dos

operadores de PG, com base em Davis (1989). Em todas as gerações ao longo da evolução, o mecanismo avalia se o operador genético está produzindo descendentes melhores (em termos da acurácia) do que a população da qual os genitores pertencem, dentro de uma janela de avaliação composta por algumas gerações que antecedem à geração na qual se faz a análise. Se o operador estiver gerando descendentes melhores que a população da qual os genitores pertencem, dentro da janela de análise, sua probabilidade de ocorrência aumentará na próxima geração; caso contrário, diminuirá. Caso um operador permaneça por um determinado número de gerações somente recebendo avaliações negativas (gerando descendentes piores), um aumento repentino em sua probabilidade de ocorrência será promovido, para que ele tenha a oportunidade de gerar descendentes novamente. A Figura 4.6 apresenta a evolução das probabilidades de ocorrência dos operadores ao longo de um experimento típico de PG com um conjunto de dados proposto por Silva (2007) – a linha verde em negrito é a probabilidade de ocorrência do operador de mutação; a linha azul em negrito é a probabilidade de ocorrência do operador de cruzamento. As outras linhas devem ser desconsideradas. O eixo horizontal representa as gerações do experimento; o eixo vertical apresenta o valor da probabilidade alcançada por cada operador ao longo dos experimentos.

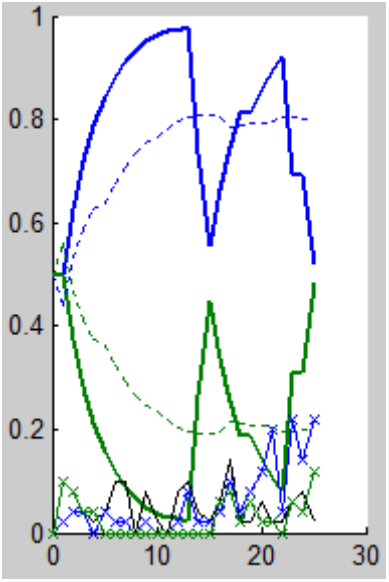

Figura 4.6 – Evolutivo das probabilidades de mutação e cruzamento em um experimento

Fonte: GPLAB (2015)

As curvas acima forma obtidas considerando-se 20 experimentos com o mesmo conjunto de dados proposto por Silva (2007). A Figura 4.7 mostra uma aproximação para a média das probabilidades de ocorrência de mutação e cruzamento para os 20 experimentos. O termo “crossover” é sinônimo de cruzamento.

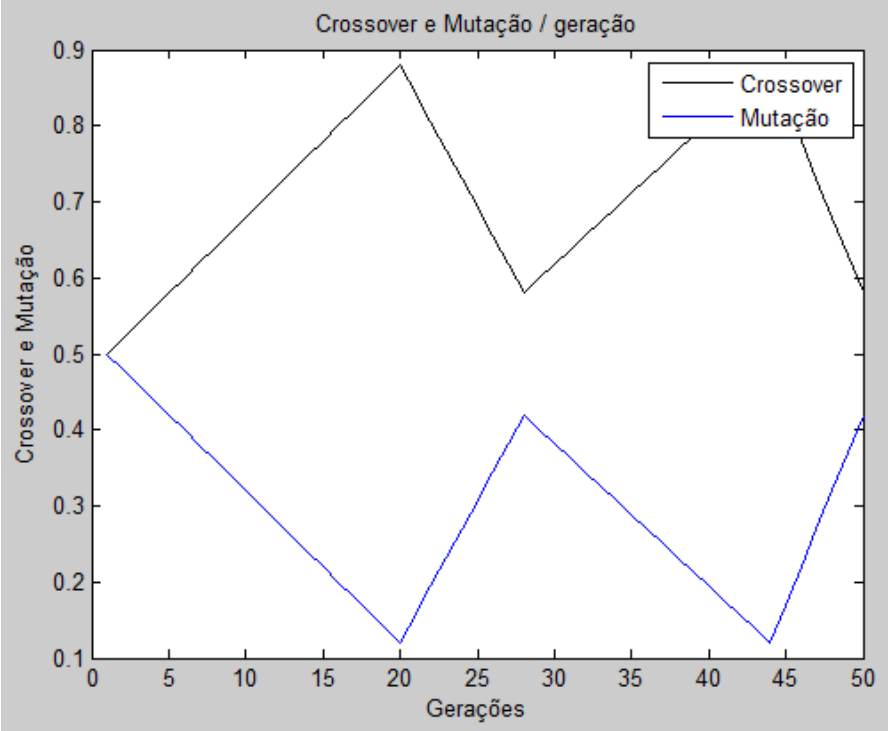

Figura 4.7 – Média das probabilidades de ocorrência de mutação e cruzamento para 20 experimentos

A curva evolutiva da Figura 4.7 foi utilizada como mecanismo de variação das probabilidades de ocorrência de mutação e cruzamento para os experimentos de PGE. Tal medida não é a mais adequada: idealmente, o mecanismo proposto por Davis (1989) e Silva (2007) deveria ser utilizado. Limitações técnicas de ajuste entre as funções do GPLAB e o GPTIPS retratam a principal razão pela qual o mecanismo proposto por Silva (2007) não foi plenamente utilizado.

Embora a média de 20 experimentos para somente um conjunto de dados retrate informação limitada com relação ao comportamento das probabilidades de ocorrência, a prática retrata uma tentativa de se propor indivíduos melhores através da maior ou menor incidência de um operador genético em distintos momentos da evolução.

4.4
Sumário

A tabela abaixo sumariza as informações do algoritmo de PGE apresentado ao longo deste capítulo.

O intervalo de cada parâmetro foi esecificado tomando por base as sugestões de Poli et al (2008): “o senso comum afirma que o ideal sobre o tamanho da população é fazê-la a maior possível, mas há alguns que sugerem que se realizem diversos experimentos com populações menores” (portanto, ao campo **Tamanho da População** foram atribuídos valores no intervalo de 50 a 150 com diversos experimentos, ao invés de utilizar valores maiores ou iguais a 500, por exemplo); “tipicamente, o número de gerações é limitado entre 10 e 50” (portanto, ao campo **Número de Gerações** foram atribuídos valores no intervalo de 15 a 100); “é comum que se crie a população inicial de forma aleatória, usando o método *ramped half-and-half*, com altura máxima no intervalo de 2 a 6” (portanto, à **Altura Máxima do Indivíduo** e à **Altura Máxima para sub-árvore criada** foram atribuídos valores no intervalo de 2 a 5).

Ao parâmetro **Número Máximo de Genes por Indivíduo** foram atribuídos valores entre 2 e 5. Experimentos preliminares mostraram que valores menores do que 2 para o parâmetro citado evidenciavam indivíduos com pouca variabilidade genética, enquanto que valores maiores do que 5 evidenciavam indivíduos com alta variabilidade, porém com pouca acurácia quando aplicados ao conjunto de validação ou teste. Portanto, considerou-se o intervalo de 2 a 5 como o mais coerente em termos de variabilidade genética e acurácia.

Aos parâmetros **Probabilidade de Ocorrência de Mutação tradicional dado que ocorrerá Mutação** e **Probabilidade de Ocorrência de Cruzamento intragênico, dado que ocorrerá Cruzamento** foram atribuídos valores *default* do software GPTIPS.

Tabela 4.2 – PGE: sumário

Objetivo	Obter modelos de regressão e classificação com maior acurácia possível.
Conjunto de Funções	Soma e Multiplicação.
Conjunto de Terminais (Ω)	Variáveis do conjunto de dados.
Acurácia	REQM (para tarefas de regressão); %_$\boldsymbol{inc}$ (para tarefas de classificação).
Seleção	Por Torneio, com variante da pressão lexicográfica de Luke & Panait (2002) com significância estatística. Tamanho do torneio: 7 indivíduos.
Parâmetros	
- Tamanho da População	De 50 a 150.
- Número de Gerações	De 15 a 100.
- Altura Máxima do Indivíduo	De 2 a 5.
- Altura Máxima para sub-árvore criada	De 2 a 5.
- Número Máximo de Genes por Indivíduo	De 2 a 5.
- Probabilidades de Ocorrência de Mutação e Cruzamento	Variantes ao longo da evolução (seguem a Figura 4.7).
- Probabilidade de Ocorrência de Mutação tradicional (Koza, 1992), dado que ocorrerá Mutação.	De 50% a 95%.
- Probabilidade de Ocorrência de Cruzamento intragênico, dado que ocorrerá Cruzamento.	50%.
- Taxa de Elitismo	5% sobre a população (fixo).
- Condição de Parada	Número de gerações atingido.

5
Experimentos e Resultados

5.1
Conjuntos de Dados

Utiliza-se a PGE, com os parâmetros descritos na seção 4.4, para gerar modelos de regressão linear e não linear para alguns conjuntos de dados, listados nas tabelas a seguir. Todos os conjuntos de dado são oriundos do UCI *Machine Learning Repository*. A descrição completa dos conjuntos de dados pode ser visualizada na informação "Sítio", em cada uma das tabelas.

Tabela 5.1 – Conjunto de Dados: Concreto

Nome do Conjunto de Dados	Resistência à Compressão do Concreto
Nome Abreviado	Concreto
Tipo de Tarefa	Regressão
Variável de Resposta	Resistência à compressão do concreto.
Breve Descrição	A resistência à compressão do concreto é uma função não linear de atributos como tempo de fabricação e ingredientes usados na fabricação, presentes no conjunto de dados.
Área do Conjunto de Dados	Engenharia / Física
Sítio	https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
Número de Exemplos (n)	1.030
Número de Atributos / Variáveis	8

Tabela 5.2 – Conjunto de Dados: Casas

Nome do Conjunto de Dados	Casas
Nome Abreviado	Casas
Tipo de Tarefa	Regressão
Variável de Resposta	Preços de casas (em US$) no subúrbio de Boston.
Breve Descrição	Propõe a estimação de valores (em US$) de casas no subúrbio de Boston, em função de variáveis relacionadas ao setor imobiliário e componentes socioeconômicas.
Área do Conjunto de Dados	Economia / Finanças
Sítio	https://archive.ics.uci.edu/ml/datasets/Housing
Número de Exemplos (n)	506
Número de Atributos / Variáveis	13

Tabela 5.3 – Conjunto de Dados: Ruídos

Nome do Conjunto de Dados	Ruídos em Aerofólios
Nome Abreviado	Ruídos
Tipo de Tarefa	Regressão
Variável de Resposta	Nível de pressão do ruído no aerofólio, em decibéis.
Breve Descrição	Séries de experimentos aerodinâmicos e testes acústicos sobre seções de aerofólio em 2 e 3 dimensões, conduzidos em túneis de vento; propõem estimar o nível de pressão do ruído no aerofólio, em decibéis.
Área do Conjunto de Dados	Engenharia / Física
Sítio	https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise
Número de Exemplos (n)	1.503
Número de Atributos / Variáveis	5

Tabela 5.4 – Conjunto de Dados: Proteínas

Nome do Conjunto de Dados	Propriedades Físico-Químicas da Estrutura Terciária de Proteínas
Nome Abreviado	Proteínas
Tipo de Tarefa	Regressão
Variável de Resposta	Tipo de propriedade da estrutura da proteína.
Breve Descrição	Não há.
Área do Conjunto de Dados	Biologia / Ciências da Natureza
Sítio	http://archive.ics.uci.edu/ml/datasets/Physicochemical+Properties+of+Protein+Tertiary+Structure
Número de Exemplos (n)	45.730
Número de Atributos / Variáveis	9

Tabela 5.5 – Conjunto de Dados: Iates

Nome do Conjunto de Dados	Estudo Hidrodinâmico de Iates
Nome Abreviado	Iates
Tipo de Tarefa	Regressão
Variável de Resposta	Resistência residual ao deslocamento.
Breve Descrição	Busca-se estimar a resistência residual ao deslocamento de iates, quando em fase inicial de construção, pois julga-se que essa variável tem efeito causal sobre o desempenho do iate e sua força propulsora.
Área do Conjunto de Dados	Física / Engenharia
Sítio	http://archive.ics.uci.edu/ml/datasets/Yacht+Hydrodynamics
Número de Exemplos (n)	308
Número de Atributos / Variáveis	6

Tabela 5.6 – Conjunto de Dados: Wisconsin

Nome do Conjunto de Dados	Câncer de Mama – Wisconsin
Nome Abreviado	Wisconsin
Tipo de Tarefa	Classificação
Variável de Resposta	Pessoa tem ou não tem câncer.
Breve Descrição	Conjunto de dados obtido do Hospital da Universidade de Wisconsin, coletado ao longo de dois anos (1989 a 1991).
Área do Conjunto de Dados	Saúde
Sítio	http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
Número de Exemplos (n)	699
Número de Atributos / Variáveis	9

Tabela 5.7 – Conjunto de Dados: Diabetes

Nome do Conjunto de Dados	Diabetes em Índios Pima
Nome Abreviado	Diabetes
Tipo de Tarefa	Classificação
Variável de Resposta	Pessoa tem ou não tem diabetes.
Breve Descrição	A variável de resposta evidencia se a paciente apresenta ou não sinais de diabetes de acordo com o critério da Organização Mundial de Saúde (OMS).
Área do Conjunto de Dados	Saúde
Sítio	http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
Número de Exemplos (n)	768
Número de Atributos / Variáveis	8

Tabela 5.8 – Conjunto de Dados: Ionosfera

Nome do Conjunto de Dados	Ionosfera
Nome Abreviado	Ionosfera
Tipo de Tarefa	Classificação
Variável de Resposta	Há ou não há estrutura conhecida para as radiações observadas.
Breve Descrição	16 antenas de alta frequência enviam radiações para a ionosfera. A reação em elétrons livres pode gerar algum tipo de estrutura – que é a variável de resposta desse conjunto de dados.
Área do Conjunto de Dados	Física
Sítio	http://archive.ics.uci.edu/ml/datasets/Ionosphere
Número de Exemplos (n)	351
Número de Atributos / Variáveis	33

5.2
Evolução das Métricas de Desempenho e Metodologia de Comparação de Modelos

Um experimento de PGE fornece uma família de modelos de regressão ou classificação, em função do tipo de conjunto de dados associado. Geração a geração, entre os indivíduos que competem entre si para proliferar seus descendentes nas próximas gerações, há um melhor indivíduo por geração, em função da métrica de acurácia utilizada – este grupo será referenciado como (grupo dos) melhores indivíduos por geração de um experimento.

Em dois ou mais experimentos independentes de PG (por consequência, da PGE), não há qualquer garantia de similaridade, tanto em forma quanto em acurácia, entre os indivíduos gerados nos distintos experimentos, população a população, devido à natureza estocástica da PG. Portanto, é usual que se realize uma série de experimentos de PG, avaliando-se não o comportamento das métricas de interesse para um único indivíduo por geração e sim para uma média de melhores indivíduos, geração a geração, tomando como referência todos os experimentos considerados para cálculo da média.

As métricas de interesse são obtidas em função da métrica de acurácia – otimizada ao longo do processo evolutivo – dos melhores indivíduos por geração. A métrica de acurácia também é uma métrica de interesse, de tal forma que os termos são intercambiados na sequência do texto.

Realizaram-se 10 experimentos para cada um dos conjuntos de dados citados e a média das métricas de interesse é calculada para os mesmos 10 experimentos. Embora tenha se realizado um número muito superior a 10 experimentos para cada conjunto de dados, a evolução das médias das métricas de interesse considera um conjunto de 10 experimentos. A razão para o uso deste número reside no fato de o Laboratório de Inteligência Computacional do Departamento de Informática da Universidade de Nicolau Copérnico manter resultados de algoritmos para classificação em diversos conjuntos de dados e utilizar a validação cruzada k-*fold*, com $\mathrm{k} = 10$, como metodologia de comparação entre os algoritmos. k, do termo k-*fold*, não possui qualquer relação com as k variáveis independentes de $X = \{x_1, x_2, \dots, x_k\}$.

A avaliação de um algoritmo por validação cruzada k-*fold* é realizada da seguinte forma (Kohavi, 1995): divide-se aleatoriamente o conjunto de dados em k subconjuntos mutuamente exclusivos, de tamanhos semelhantes; avalia-se o algoritmo nos conjuntos de treino e teste, sendo que um dos k subconjuntos é selecionado para ser o conjunto de teste e os $k-1$ restantes constituem o conjunto de treino; o algoritmo será avaliado k vezes em conjuntos de treino e teste, sendo, a cada avaliação, um dos k subconjuntos o conjunto de teste e os $k-1$ subconjuntos restantes o conjunto de treino. Na validação cruzada k-*fold*, cada um dos k conjuntos de teste recebe o nome de conjunto de validação.

Portanto, ao realizar-se a avaliação dos algoritmos pela metodologia proposta pelo laboratório citado, obtém-se automaticamente um grupo de *benchmarks* para cada conjunto de dados associado somente à tarefa de classificação, pois o referido laboratório não disponibiliza o mesmo acervo para algoritmos e conjuntos de dados associados à regressão.

Para efeito de uniformização das metodologias de avaliação e comparação de modelos, são avaliados e comparados modelos associados aos conjuntos de dados de regressão e classificação por meio da validação cruzada 10-*fold*.

Como visto anteriormente, sob a ótica dos modelos de regressão, são de interesse ao longo da evolução as métricas de REQM e $\overline{R}^2$. Sob a ótica dos modelos de classificação, é de interesse a métrica $\%\_inc$. Como a PGE tem natureza parcimoniosa, é coerente que se observe o comportamento do número de regressores, estatisticamente significantes ou não, ao longo da evolução.

O laboratório citado ordena os algoritmos em cada conjunto de dados pela média de desempenho $1-\%\_inc$ (o percentual de classificações corretas) no conjunto de validação para a validação cruzada 10-*fold*. Logo, a grandeza $1-\%\_inc$ também será representada nos gráficos evolutivos de métricas médias de desempenho.

Há somente uma diferença entre as metodologias de avaliação e comparação de modelos para regressão e classificação: enquanto para os modelos de classificação a metodologia segue estritamente a validação cruzada 10-*fold* proposta pelo laboratório, para os modelos de regressão divide-se o conjunto de dados inicialmente em dois grupos – treino, com 70% do total de exemplos do

conjunto de dados, e teste, com o restante – para, posteriormente, realizar validação cruzada 10-*fold* no conjunto de treino.

Os gráficos das Figuras 5.1 a 5.8, que mostram a evolução das métricas de interesse para os distintos conjuntos de dados, apresentam abreviações ou outras formas de representação de algumas grandezas, são elas: “R2 Ajustado” para $\overline{R}^2$; “#reg” para número de regressores; “#reg-ES” para número de regressores estatisticamente significantes; “%-inc” para $\%\_inc$; “%-corr” para $1 - \%\_inc$.

Os gráficos evolutivos das métricas de interesse foram fornecidos por experimentos de PGE com os parâmetros listados na tabela 5.9 – os valores dos parâmetros estão dentro do intervalo apresentado para cada um deles na tabela 4.2.

Tabela 5.9 – Parâmetros de um experimento de PGE

Parâmetros	
- Tamanho da População	150
- Número de Gerações	25 (regressão); 15 (classificação).
- Altura Máxima do Indivíduo	3
- Altura Máxima para sub-árvore criada	3
- Número Máximo de Genes por Indivíduo	3
- Probabilidades de Ocorrência de Mutação e Cruzamento	Variantes ao longo da evolução (seguem a Figura 4.7).
- Probabilidade de Ocorrência de Mutação tradicional (Koza 1992), dado que ocorrerá Mutação.	50%.
- Probabilidade de Ocorrência de Cruzamento intragênico, dado que ocorrerá Cruzamento.	50%.
- Taxa de Elitismo	5% sobre a população (fixo).
- Condição de Parada	Número de gerações atingido.

A razão de escolha dos valores para cada parâmetro, listados na tabela 5.9, reside no fato de experimentos iniciais com a PGE mostrarem que o **Tamanho da População** fixado em 150, o **Número de Gerações** fixado entre os valores de 15 à 25, **Altura Máxima do Indivíduo** e **Altura Máxima para sub-árvore criada** e **Número Máximo de Genes por Indivíduo** fixados no valor 3, a **Probabilidade de Ocorrência de Mutação tradicional dado que ocorrerá Mutação** e **Probabilidade de Ocorrência de Cruzamento intragênico dado que ocorrerá Cruzamento** fixados em 50% geravam grande variabilidade genética com acurácia elevada frente a outros experimentos com os parâmetros assumindo valores distintos. Não será apresentado um estudo de sensibilidade a respeito desses parâmetros na PGE, pois não é objetivo da dissertação. Reconhece-se que sejam necessários esforços adicionais para determinar um conjunto de parâmetros mais adequado – via análise de sensibilidade – ou ótimo – via implementação de algoritmo genético – para a PGE e que tais conjuntos de parâmetros podem ser ainda função do conjunto de dados aos quais a PGE é aplicada.

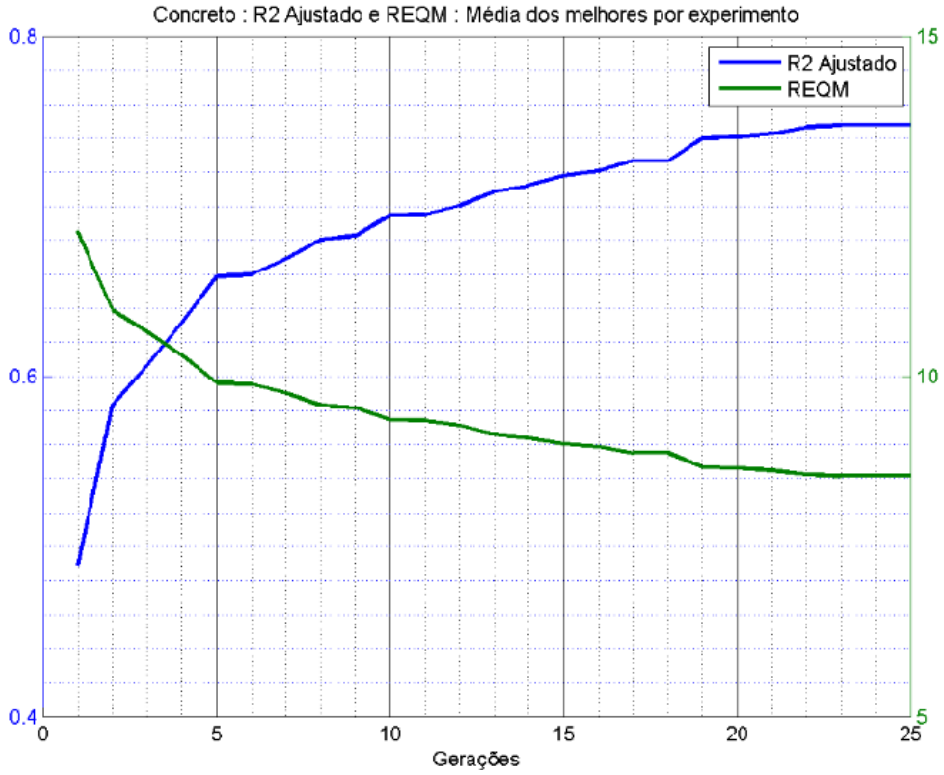

Figura 5.1a – Concreto: $\bar{\mathbf{R}}^{\mathbf{2}}$ e REQM

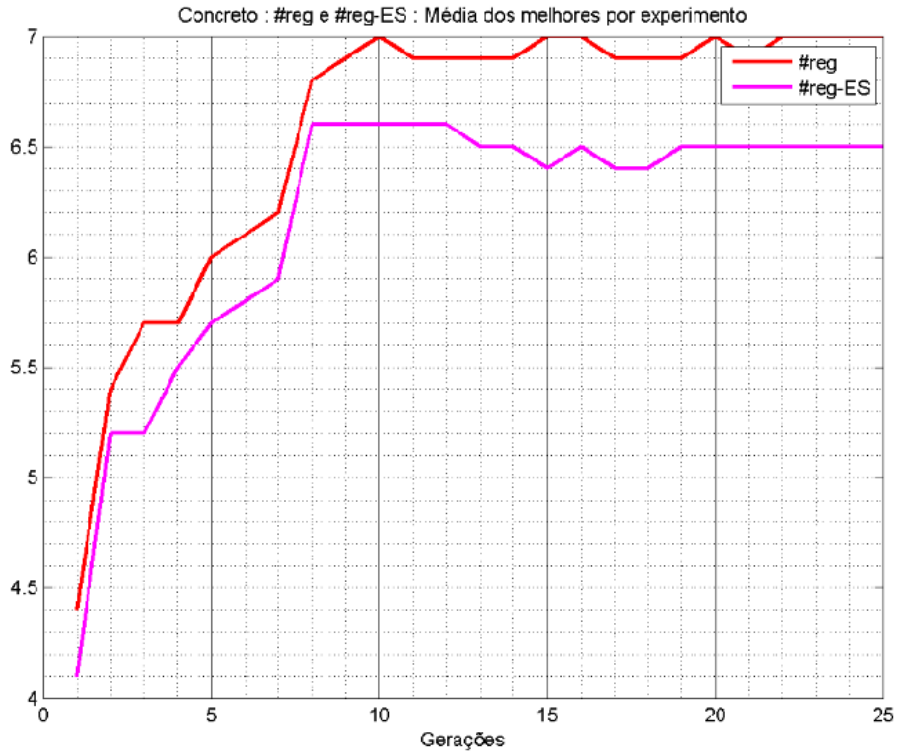

Figura 5.1b – Concreto: #reg e #reg-ES

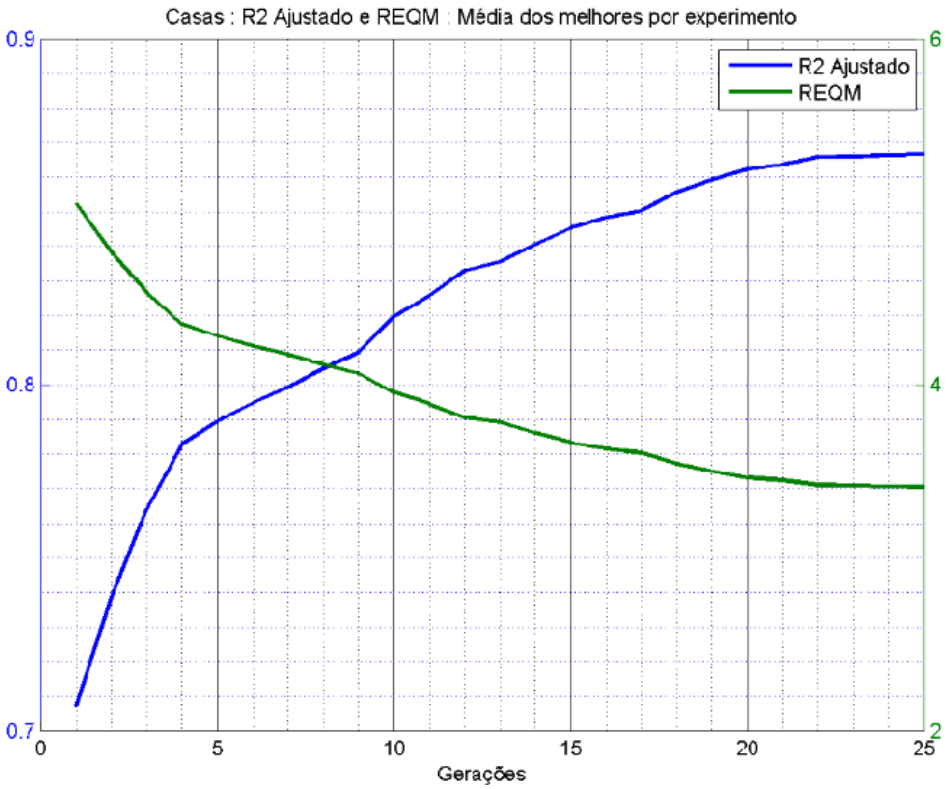

Figura 5.2a – Casas: $\overline{\mathbf{R}}^2$ e REQM

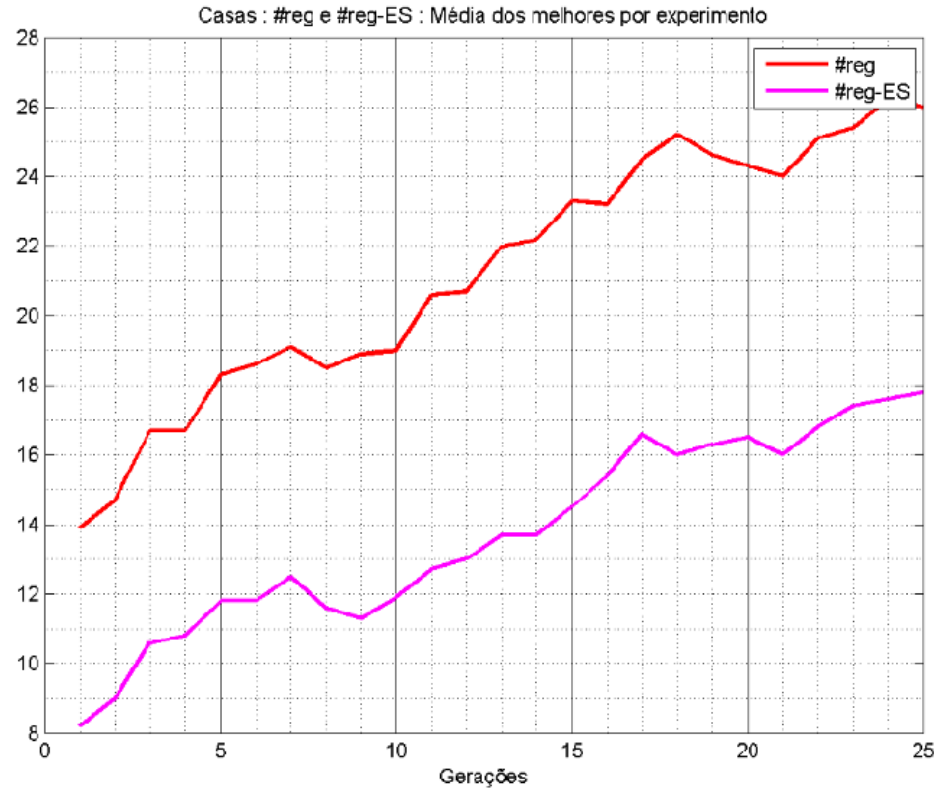

Figura 5.2b – Casas: #reg e #reg-ES

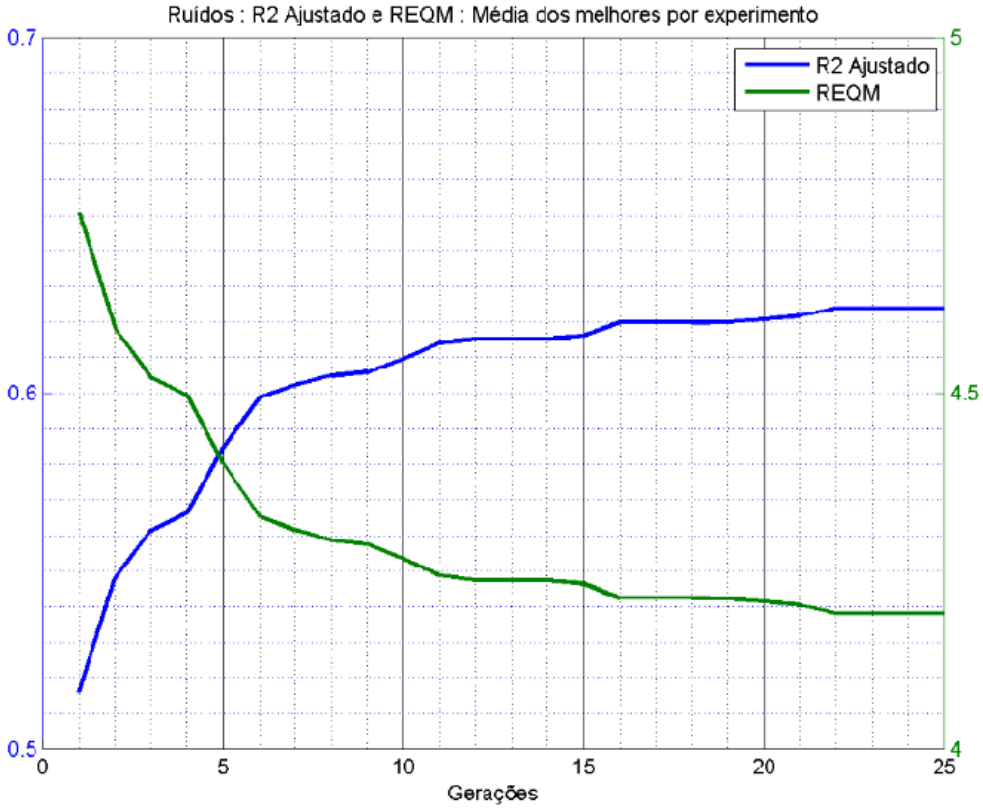

Figura 5.3a – Ruídos: $\overline{\mathbf{R}}^{\mathbf{2}}$ e REQM

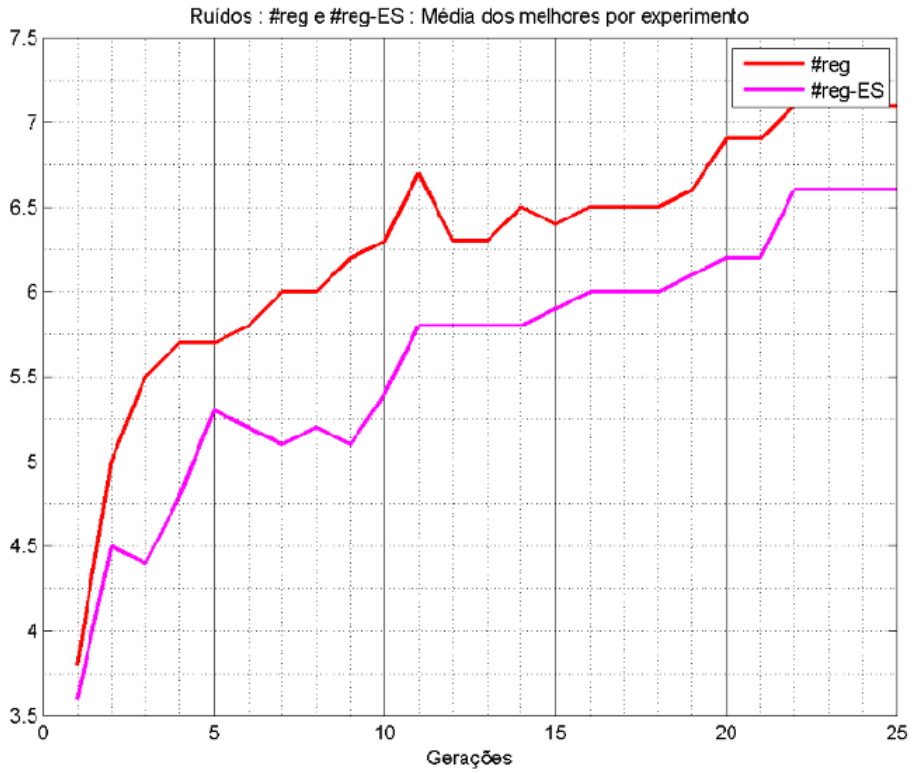

Figura 5.3b – Ruídos: #reg e #reg-ES

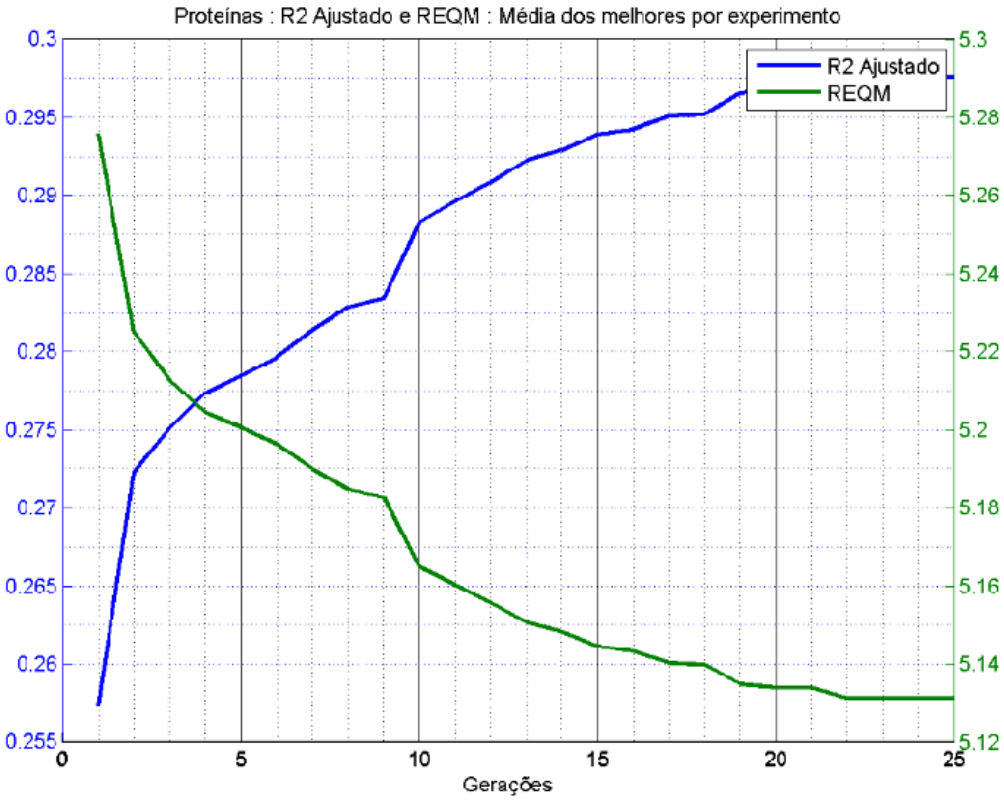

Figura 5.4a – Proteínas: $\overline{\mathbf{R}}^2$ e REQM

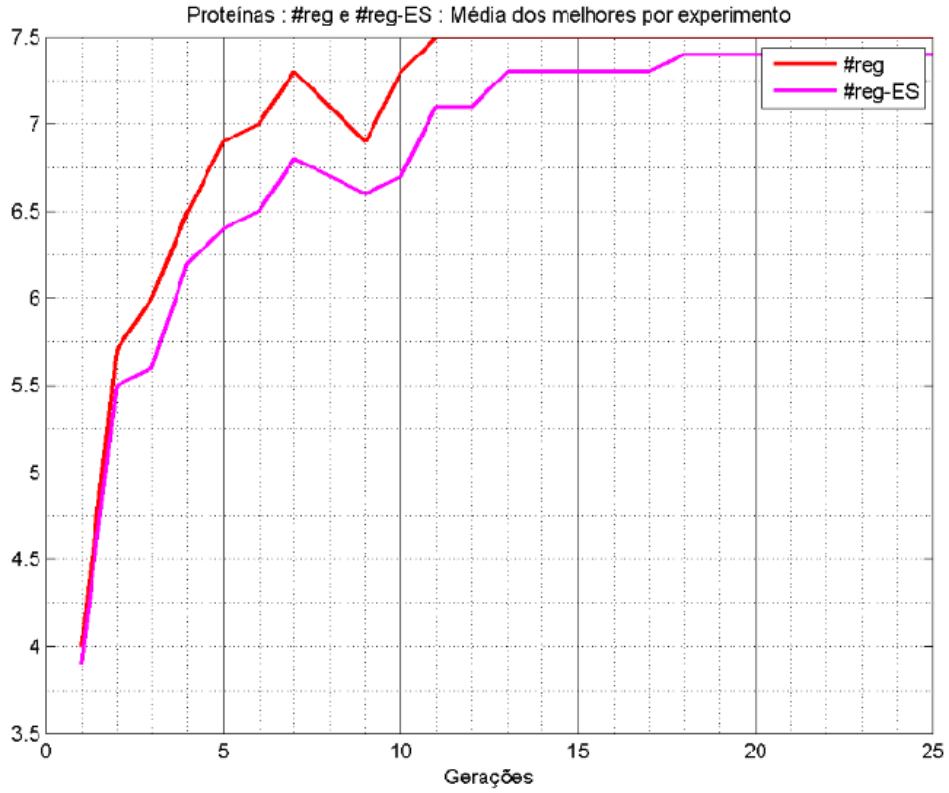

Figura 5.4b – Proteínas: #reg e #reg-ES

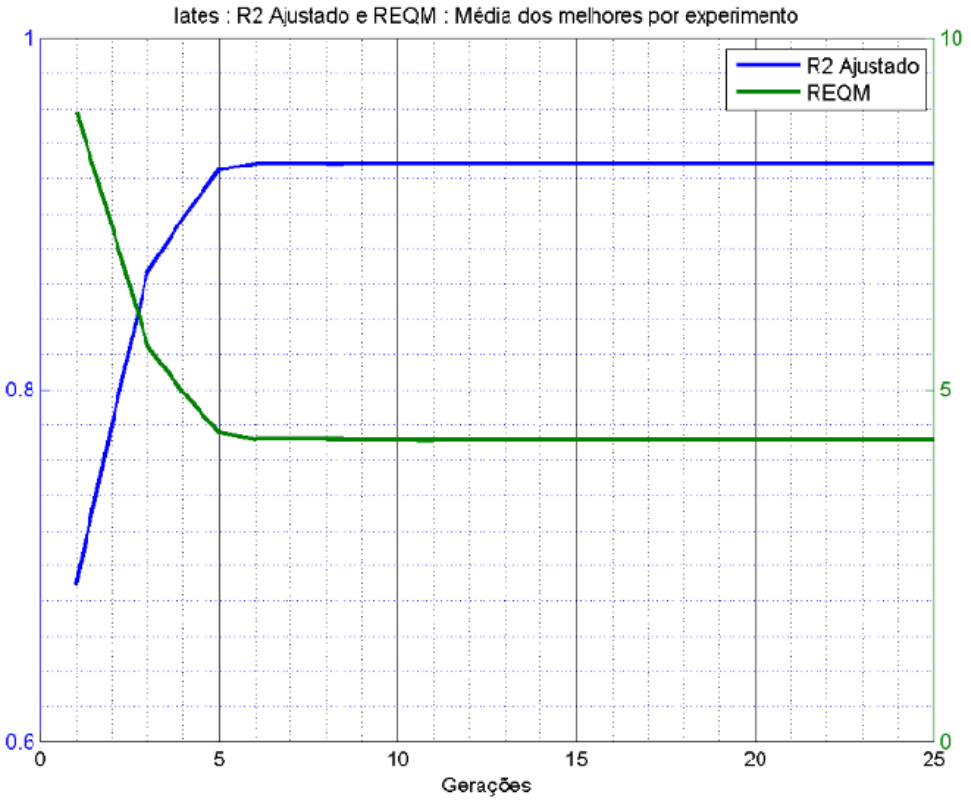

Figura 5.5a – Iates: $\overline{\mathbf{R}}^{\mathbf{2}}$ e REQM

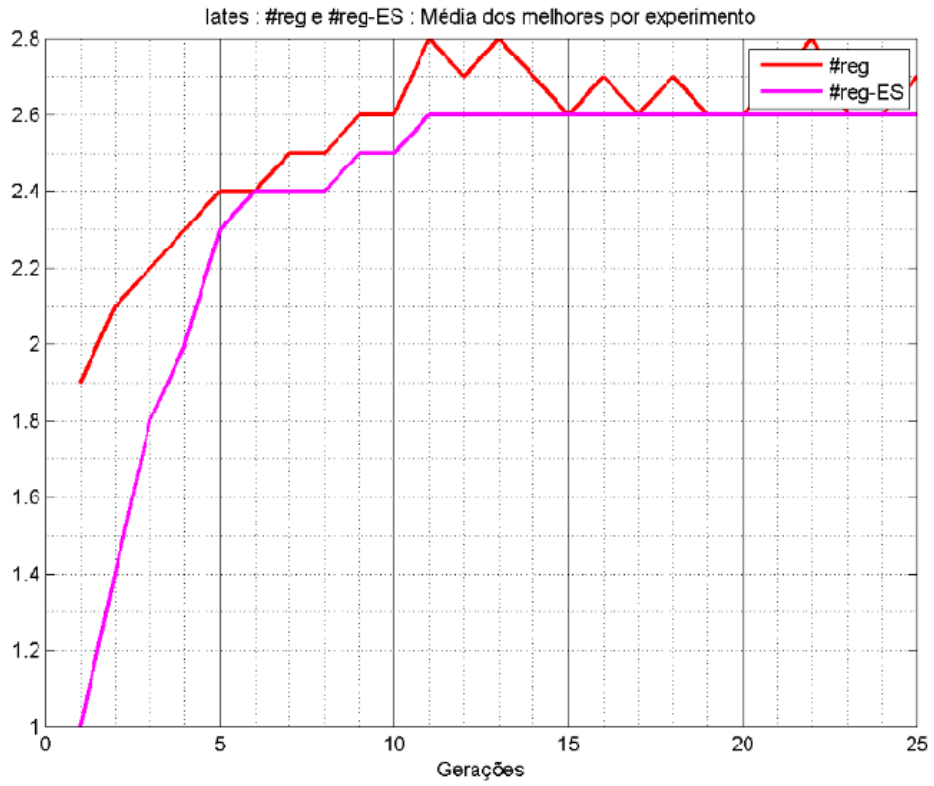

Figura 5.5b – Iates: #reg e #reg-ES

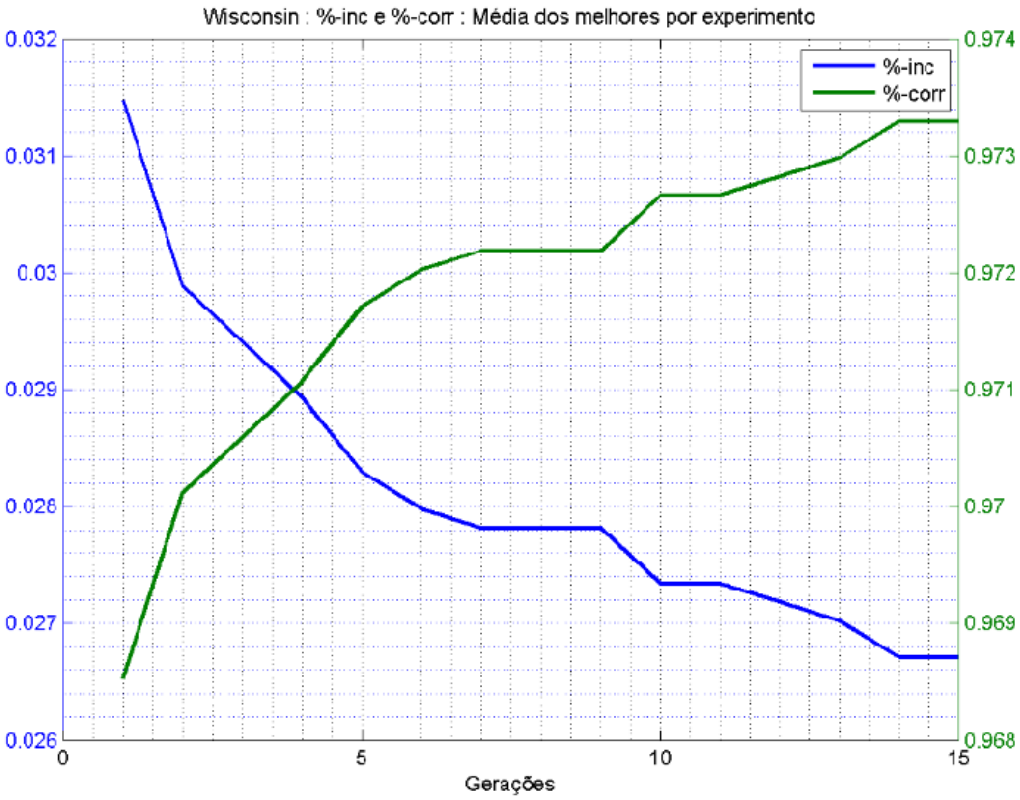

Figura 5.6a – Wisconsin: %-inc e %-corr

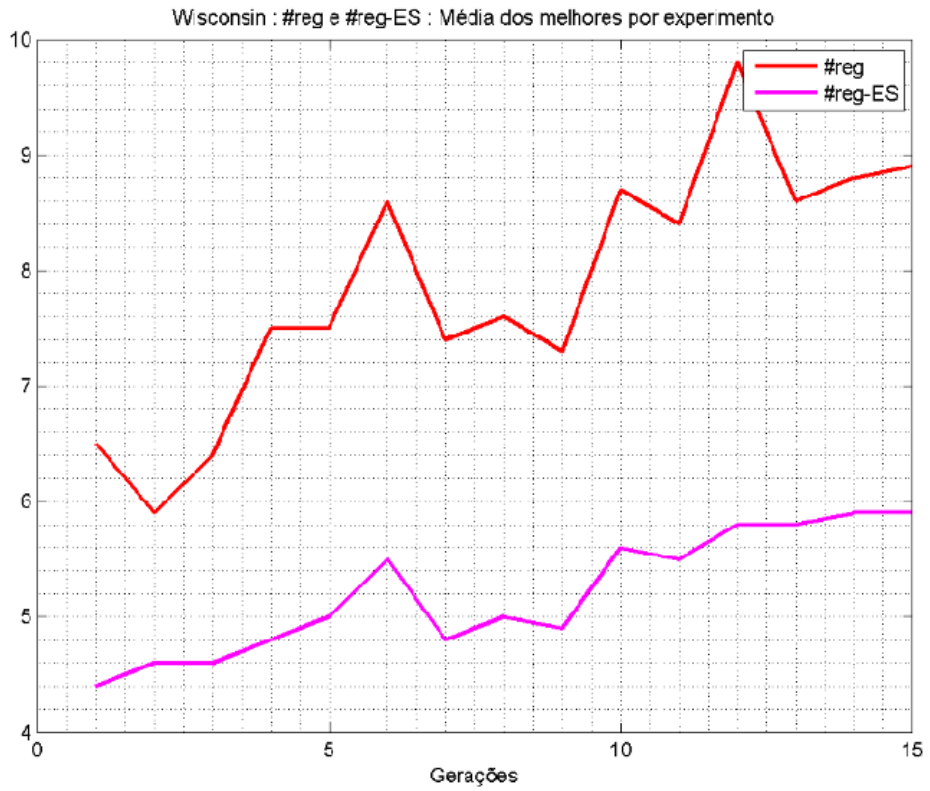

Figura 5.6b – Wisconsin: #reg e #reg-ES

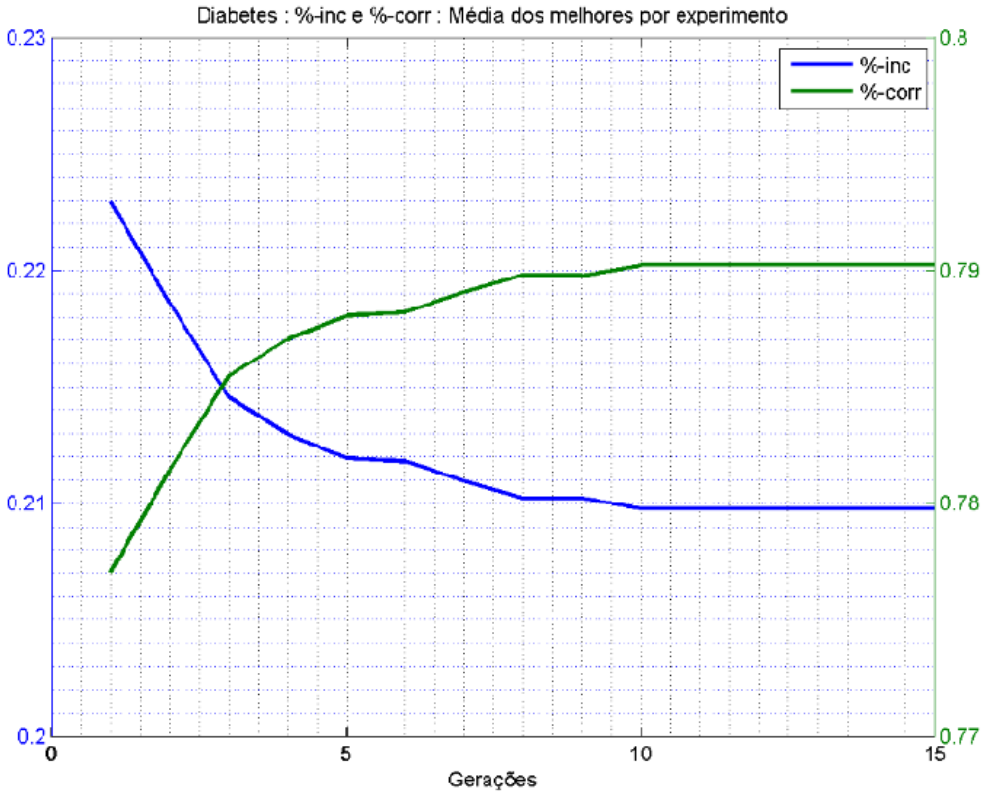

Figura 5.7a – Diabetes: %-inc e %-corr

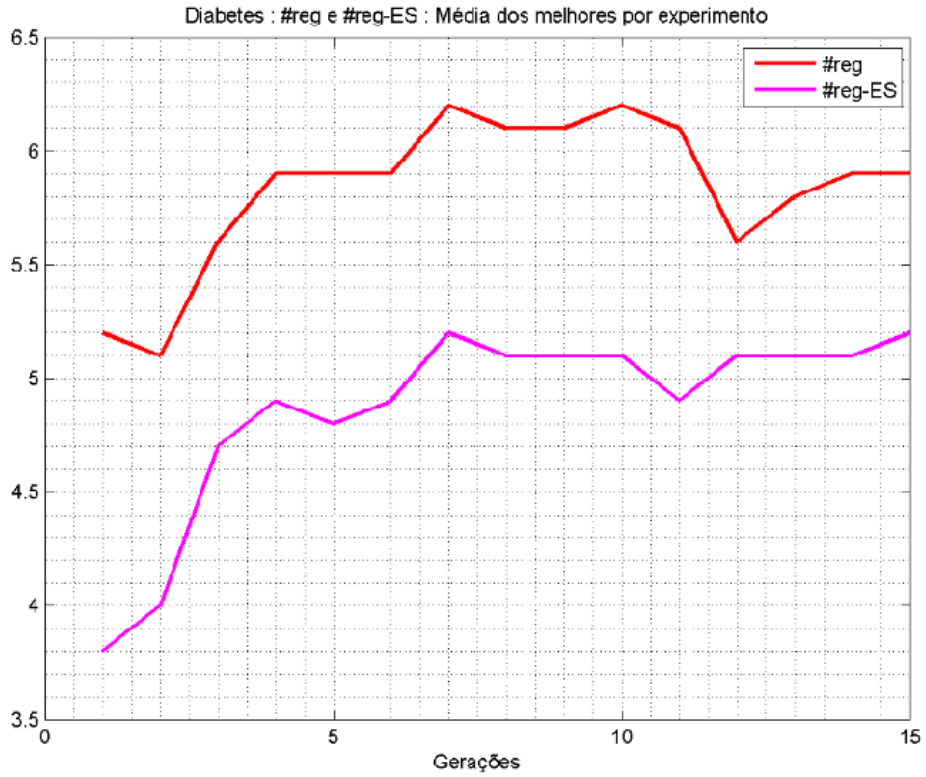

Figura 5.7b – Diabetes: #reg e #reg-ES

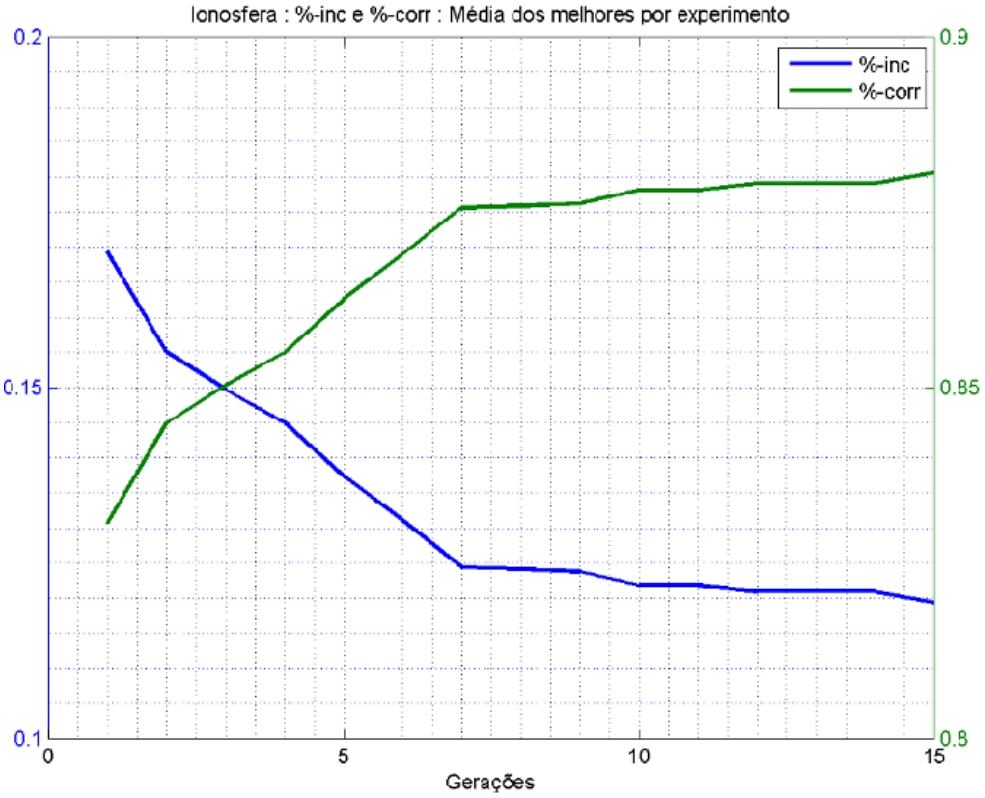

Figura 5.8a – Ionosfera: %-inc e %-corr

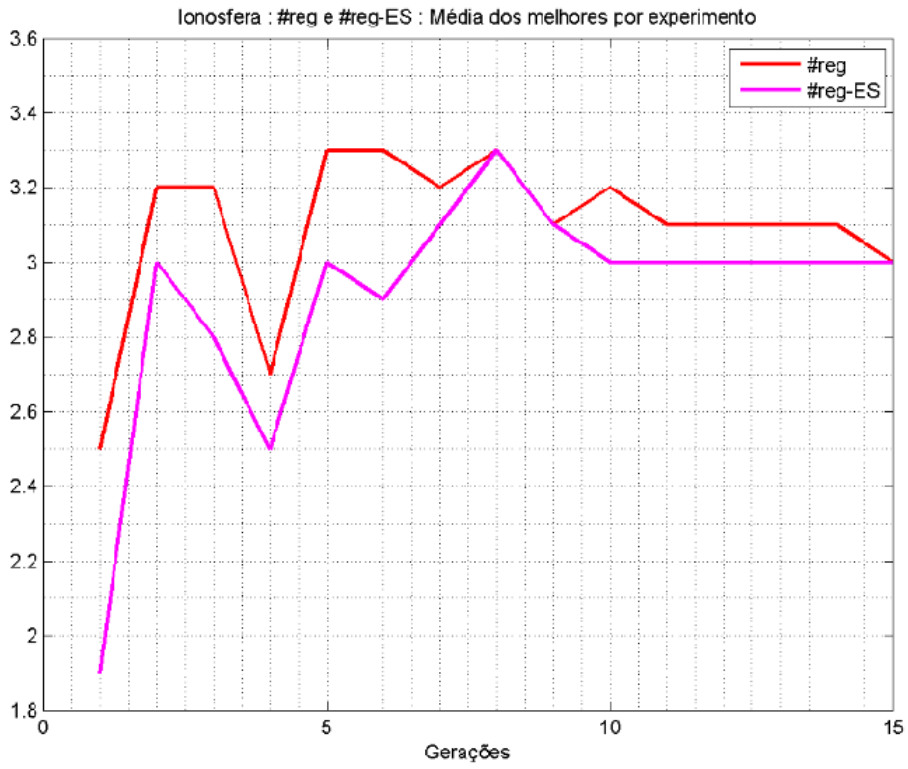

Figura 5.8b – Ionosfera: #reg e #reg-ES

5.3
Benchmarks

O grupo de *benchmarks* para os modelos de classificação foi definido na seção 5.2.

Como determinado anteriormente, o algoritmo de geração de modelos de regressão linear se utiliza da prova matemática relacionada ao acréscimo de x_{k+1} a X (o acréscimo de regressores nunca diminui o R^2, consequentemente nunca aumenta o REQM) e da condição necessária para que x_{k+1} seja estatisticamente significante e, por consequência, gerador de aumento a $\overline{R}^2$.

É possível propor um *benchmark* para os algoritmos de regressão da PGE a partir do resultado teórico acima e da análise dos gráficos de evolução do número de regressores. Caso a PG não fosse utilizada como mecanismo de geração de modelos, o simples acréscimo de regressores a X segundo uma regra que não explora o espaço de busca seria suficiente. A rotina $x2fx$, do *Matlab*, realiza essa tarefa. Será dado um exemplo para pleno entendimento de seu funcionamento. Supondo x_1, x_2, x_3 as variáveis de Ω, $x2fx$ gerará o conjunto

$X = \{x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3,\ x_1^2,\ x_2^2,\ x_3^2\}$ de regressores, composta pelas variáveis independentes originais de Ω, seus termos cruzados e termos quadráticos. É possível utilizar $x2fx$ novamente, agora sobre $x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3,\ x_1^2,\ x_2^2$ e x_3^2, para que se gere um novo conjunto X com ainda mais regressores. No caso citado acima, há a formação de dois modelos para comparação com a PGE: o primeiro deles após aplicação única da rotina $x2fx$ e o segundo após a aplicação dupla da rotina $x2fx$ (a segunda aplicação é realizada sobre o conjunto de variáveis originadas da primeira aplicação de $x2fx$).

Portanto, os *benchmarks* dos modelos de regressão linear gerados pela PGE serão os dois modelos de regressão linear criados sob utilização da rotina $x2fx$, nomeados "*Benchmark* 1" e "*Benchmark* 2", de tal forma que o primeiro deles é o que possui X com menor cardinalidade.

5.4
Resultados

Esta seção apresenta os resultados comparativos entre os algoritmos.

5.4.1
Regressão

Cada uma das Figuras, de 5.9 à 5.13, apresenta um conjunto de resultados para um conjunto de dados. Há quatro quadros compondo cada uma das Figuras. Há um quadro relativo ao $\overline{R}^2$ no conjunto de treino para os três algoritmos testados, em cada um dos k experimentos – este quadro tem o título "treino". "Exp" é a abreviação de experimento; "Bench 1" é a abreviação de *Benchmark* 1; "Bench 2" é a abreviação de *Benchmark* 2. "Média", a métrica mais importante de cada quadro, realiza a média da grandeza mostrada no quadro considerando todos os experimentos, para cada algoritmo. "DP" é o desvio padrão amostral da grandeza mostrada no quadro considerando todos os experimentos, para cada algoritmo.

Há um quadro relativo a $\overline{R}^2$ no conjunto de validação para os três algoritmos testados, em cada um dos k experimentos – este quadro tem o título

"validação". Por fim, há um quadro relativo a $\overline{R}^2$ no conjunto de teste para os três algoritmos testados, em cada um dos k experimentos – este quadro tem o título "teste". O quadro com título "#reg-ES" é semelhante aos quadros anteriores, com a única diferença de representar o número de regressores estatisticamente significantes.

Médias em azul apontam o algoritmo que obteve o melhor desempenho por quadro: quanto maior o $\overline{R}^2$ atingido, melhor o resultado do algoritmo. Quanto menor o número de regressores, mais parcimonioso é o modelo e melhor é o resultado do algoritmo (supondo que apresente uma boa métrica de $\overline{R}^2$).

Concreto

	Treino		
Exp	PGE	Bench 1	Bench 2
1	0,7474	0,6761	0,2588
2	0,7674	0,5974	0,4489
3	0,7495	0,5631	0,1900
4	0,7316	0,6931	0,2997
5	0,7488	0,5791	0,4127
6	0,7569	0,5651	0,5875
7	0,7421	0,5790	0,5019
8	0,7453	0,6802	0,2264
9	0,7444	0,5908	0,2247
10	0,7450	0,5720	0,5501
Média:	0,7478	0,6096	0,3701
DP:	0,0094	0,0520	0,1478

	Validação		
Exp	PGE	Bench 1	Bench 2
1	0,6137	0,5565	0,3392
2	0,8826	0,4245	0,5496
3	0,8726	0,2630	0,2947
4	0,8026	0,7186	-0,1029
5	0,6741	0,3326	0,3012
6	0,7981	0,1849	0,3170
7	0,8832	0,4527	0,5428
8	0,5562	0,5368	0,0981
9	0,7264	0,6655	-0,0006
10	0,6698	0,2230	0,3174
Média:	0,7479	0,4358	0,2657
DP:	0,1175	0,1846	0,2122

	Teste		
Exp	PGE	Bench 1	Bench 2
1	0,7528	0,7809	0,2514
2	0,7753	0,7254	0,4076
3	0,7860	0,6703	0,1946
4	0,7776	0,8251	0,2286
5	0,7609	0,7101	0,3919
6	0,7853	0,6716	0,6022
7	0,7823	0,7096	0,3790
8	0,7492	0,7897	0,1748
9	0,7894	0,7142	0,1905
10	0,7780	0,6843	0,5846
Média:	0,7737	0,7281	0,3405
DP:	0,0143	0,0530	0,1593

	#reg-ES		
Exp	PGE	Bench 1	Bench 2
1	6	22	11
2	7	32	15
3	6	32	3
4	6	27	13
5	6	29	17
6	7	24	24
7	7	28	15
8	6	27	10
9	7	28	8
10	7	31	17
Média:	6,50	28,00	13,30
DP:	0,53	3,27	5,76

Figura 5.9 – Resultados para o conjunto de dados Concreto

Casas

	Treino		
Exp	PGE	Bench 1	Bench 2
1	0,8690	0,7552	NaN
2	0,8603	0,7594	NaN
3	0,8579	0,7396	NaN
4	0,8799	0,7400	NaN
5	0,8808	0,7401	NaN
6	0,8571	0,8143	NaN
7	0,8872	0,7461	NaN
8	0,8619	0,8211	NaN
9	0,8615	0,7668	NaN
10	0,8514	0,8528	NaN
Média:	0,8667	0,7735	#DIV/0!
DP:	0,0120	0,0407	#DIV/0!

	Validação		
Exp	PGE	Bench 1	Bench 2
1	1,3823	0,7228	-3715,2
2	1,4308	0,7098	-338,4
3	1,9355	1,1575	-22,0
4	2,4634	1,2671	-268,3
5	-9,8289	0,2852	-16,6
6	0,6060	0,5306	-57,9
7	0,0100	0,6184	-21,9
8	1,0717	0,2517	-25,2
9	1,1945	1,1423	-23,3
10	0,2703	-0,1477	-242,2
Média:	0,0536	0,6538	-473,1128
DP:	3,5491	0,4513	1145,7977

	Teste		
Exp	PGE	Bench 1	Bench 2
1	0,9997	0,8260	-1983,7
2	0,9487	0,7798	-3667,8
3	1,0436	0,7971	-942,1
4	0,9257	0,8132	-17313,4
5	0,9873	0,7676	-810,4
6	0,9716	0,9987	-419,8
7	0,9505	0,7334	-3101,2
8	0,9499	0,8950	-646,8
9	0,9935	0,8064	-392,2
10	0,8182	0,8628	-3424,9
Média:	0,9589	0,8280	-3270,2280
DP:	0,0598	0,0756	5097,2618

	#reg-ES		
Exp	PGE	Bench 1	Bench 2
1	19	19	319
2	17	15	318
3	14	15	319
4	21	15	318
5	20	13	319
6	17	18	319
7	23	16	318
8	18	23	319
9	16	21	318
10	13	23	319
Média:	17,80	17,80	318,60
DP:	3,08	3,58	0,52

Figura 5.10 – Resultados para o conjunto de dados Casas

Ruídos

	Treino		
Exp	PGE	Bench 1	Bench 2
1	0,6261	0,5875	0,4842
2	0,6078	0,5500	0,4845
3	0,6320	0,5884	0,4833
4	0,6081	0,6121	0,4826
5	0,6296	0,5805	0,4813
6	0,6163	0,6036	0,4836
7	0,6295	0,5846	0,4847
8	0,6416	0,5874	0,4860
9	0,6206	0,6174	0,4869
10	0,6270	0,6174	0,4835
Média:	0,6239	0,5929	0,4841
DP:	0,0107	0,0206	0,0016

	Validação		
Exp	PGE	Bench 1	Bench 2
1	0,4028	0,3648	10,9192
2	0,8406	0,7874	8,7376
3	0,3272	0,2550	11,2194
4	0,6382	0,6998	10,0626
5	0,5884	0,5008	9,8020
6	0,5446	0,5049	13,0006
7	0,5977	0,5691	7,8773
8	0,6096	0,5827	18,4811
9	0,6748	0,6815	27,7858
10	0,6310	0,5832	10,8646
Média:	0,5855	0,5529	12,8750
DP:	0,1416	0,1579	5,9984

	Teste		
Exp	PGE	Bench 1	Bench 2
1	0,6169	0,5879	8,3843
2	0,6562	0,5950	8,5641
3	0,6102	0,5736	7,9501
4	0,6454	0,6411	8,4311
5	0,6365	0,6100	9,1988
6	0,6206	0,6178	8,3761
7	0,6369	0,6026	8,4168
8	0,6490	0,6221	8,4023
9	0,6548	0,6447	11,6442
10	0,6534	0,6390	8,3599
Média:	0,6380	0,6134	8,7728
DP:	0,0169	0,0240	1,0544

	#reg-ES		
Exp	PGE	Bench 1	Bench 2
1	7	10	16
2	6	7	15
3	7	9	14
4	6	9	15
5	7	8	18
6	7	10	16
7	7	9	17
8	8	8	15
9	5	11	14
10	6	10	16
Média:	6,60	9,10	15,60
DP:	0,84	1,20	1,26

Figura 5.11 – Resultados para o conjunto de dados Ruídos

Proteínas

	Treino		
Exp	PGE	Bench 1	Bench 2
1	0,2995	0,3563	0,3146
2	0,3009	0,3594	0,3169
3	0,2994	0,3591	0,3213
4	0,3062	0,3567	0,3189
5	0,2988	0,3563	0,3174
6	0,3006	0,3577	0,3166
7	0,2922	0,3578	0,3188
8	0,2804	0,3548	0,3164
9	0,2936	0,3565	0,3222
10	0,3040	0,3566	0,3153
Média:	**0,2976**	**0,3571**	**0,3178**
DP:	**0,0073**	**0,0014**	**0,0025**

	Validação		
Exp	PGE	Bench 1	Bench 2
1	0,2932	0,3479	0,3471
2	0,3057	0,3574	0,3800
3	0,2960	0,3608	0,4351
4	0,2906	0,3335	0,3449
5	0,2763	0,3375	0,3638
6	0,3004	0,3553	0,4184
7	0,3020	0,3652	0,3592
8	0,2888	0,3667	0,3480
9	0,3109	0,3615	0,8114
10	0,3042	0,3559	0,3847
Média:	**0,2968**	**0,3542**	**0,4193**
DP:	**0,0100**	**0,0112**	**0,1411**

	Teste		
Exp	PGE	Bench 1	Bench 2
1	0,2963	0,3649	0,6098
2	0,3020	0,3679	0,6641
3	0,3004	0,3698	0,5609
4	0,3071	0,3639	0,5045
5	0,2955	0,3633	0,6812
6	0,3015	0,3672	0,5832
7	0,2943	0,3687	0,6241
8	0,2812	0,3659	0,6416
9	0,2958	0,3667	0,5729
10	0,3073	0,3665	0,4951
Média:	**0,2981**	**0,3665**	**0,5937**
DP:	**0,0075**	**0,0021**	**0,0627**

	#reg-ES		
Exp	PGE	Bench 1	Bench 2
1	8	33	49
2	6	31	47
3	8	34	47
4	8	32	47
5	8	35	49
6	8	35	49
7	6	32	49
8	6	30	49
9	8	34	47
10	8	34	45
Média:	**7,40**	**33,00**	**47,80**
DP:	**0,97**	**1,70**	**1,40**

Figura 5.12 – Resultados para o conjunto de dados Proteínas

lates

Exp	Treino PGE	Treino Bench 1	Treino Bench 2
1	0,9281	0,7722	0,9807
2	0,9308	0,7759	0,9963
3	0,9288	0,7779	0,9778
4	0,9288	0,7714	0,9815
5	0,9308	0,7851	0,9792
6	0,9277	0,7752	0,9799
7	0,9293	0,7767	0,9809
8	0,9280	0,9266	0,9303
9	0,9282	0,7748	0,9794
10	0,9294	0,7716	0,9798
Média:	**0,9290**	**0,7907**	**0,9766**
DP:	**0,0011**	**0,0479**	**0,0171**

Exp	Validação PGE	Validação Bench 1	Validação Bench 2
1	1,1554	0,9155	1,0919
2	1,0823	0,9728	1,0231
3	1,0536	0,9104	1,0219
4	0,8104	0,6546	0,5100
5	1,2260	1,4475	1,0754
6	0,7434	0,4967	0,8234
7	0,9199	0,7551	0,9801
8	0,6480	0,6519	0,6934
9	0,9216	0,7625	0,8922
10	1,3955	1,1876	1,0721
Média:	**0,9956**	**0,8755**	**0,9183**
DP:	**0,2308**	**0,2802**	**0,1918**

Exp	Teste PGE	Teste Bench 1	Teste Bench 2
1	1,2589	1,1780	1,1495
2	1,2568	1,1937	1,0932
3	1,2617	1,1934	1,1499
4	1,2235	1,1387	1,1376
5	1,2614	1,2172	1,1583
6	1,2203	1,1260	1,1295
7	1,2423	1,1714	1,1452
8	1,1870	1,1790	1,1807
9	1,2442	1,1695	1,1501
10	1,2804	1,2073	1,1384
Média:	**1,2436**	**1,1774**	**1,1432**
DP:	**0,0270**	**0,0284**	**0,0224**

Exp	#reg-ES PGE	#reg-ES Bench 1	#reg-ES Bench 2
1	3	1	5
2	3	1	40
3	2	1	2
4	2	1	18
5	2	1	3
6	3	1	6
7	3	1	5
8	3	2	3
9	3	1	5
10	2	1	4
Média:	**2,60**	**1,10**	**9,10**
DP:	**0,52**	**0,32**	**11,76**

Figura 5.13 – Resultados para o conjunto de dados lates

No conjunto de dados Concreto, a PGE apresentou o $\overline{R}^2$ médio mais elevado, dentre todos os algoritmos aplicados, nos conjuntos de treino, validação e teste, além de ser o modelo com X de menor cardinalidade ("#reg-ES" médio é o menor dentre todos os algoritmos). O desvio padrão para todas as métricas da PGE foi o menor entre os algoritmos. O desempenho no conjunto de teste é o de maior importância, pois representará a capacidade do algoritmo inferir o comportamento de y em um conjunto de dados no qual não utilizou para treinamento.

No conjunto de dados Casas, a PGE apresentou o $\overline{R}^2$ médio mais elevado, dentre todos os algoritmos aplicados, nos conjuntos de treino e teste. Os modelos com X de menor cardinalidade são a PGE e *Benchmark* 1, com média de

regressores estatisticamente significantes atingindo o valor de 17,80. Tal fato evidencia a superioridade na qualidade de regressores gerados pela PGE em relação ao *Benchmark* 1 neste conjunto de dados, visto que, embora tenham cardinalidade de X em igual número, a PGE apresentou melhor desempenho no conjunto de teste. O termo "NaN" indica que o $\overline{R}^2$ dos modelos gerados por *Benchmark* 2 foram expostos à situação em que $k \cong n$, com k elevado (média de 318,60) e n ligeiramente superior a k, fazendo com que $\overline{R}^2$ atinja valores negativamente muito elevados, excedendo à precisão computacional do *software* que fez o seu cálculo, como confirmam as tabelas de validação e teste.

No conjunto de dados Ruídos, *Benchmark* 2 explorou de maneira mais eficiente o espaço de busca de regressores, por consequência de modelos, do que a PGE e *Benchmark* 1 – observa-se pela cardinalidade média de X em *Benchmark* 2, quando comparada com os outros dois algoritmos. Por ser um conjunto de dados com poucas variáveis de entrada – somente cinco – é possível que seja necessário aumentar a capacidade de exploração do espaço de busca pela PGE, através da modificação de parâmetros do experimento, como aumento do número máximo de gerações, do número de genes máximo permitido e do tamanho de árvore máxima para os indivíduos.

A análise para o conjunto de dados Proteínas é semelhante a Ruídos.

No conjunto de dados Iates, a PGE apresentou o $\overline{R}^2$ médio mais elevado, dentre todos os algoritmos aplicados, nos conjuntos de validação e teste.

5.4.2 Classificação

Os algoritmos das tabelas 5.10 a 5.12, para cada um dos conjuntos de dados, são ordenados pela média de desempenho $1 - \%\_inc$ (o percentual de classificações corretas, representado na tabela por "%-corr") no conjunto de validação para a validação cruzada 10-*fold*. Para os experimentos de classificação, não há conjunto de teste.

Os resultados dos algoritmos listados nas tabelas abaixo, para cada conjunto de dados, à exceção do indicado por "PGE", são de responsabilidade do Laboratório de Inteligência Computacional do Departamento de Informática da

Universidade de Nicolau Copérnico. A coluna “Referência” é de domínio do laboratório e não necessariamente faz referência a artigos que tenham sido publicados com os resultados apresentados. As tabelas abaixo foram copiadas do site http://www.is.umk.pl/projects/datasets.html, também citado no capítulo de Referências Bibliográficas, e a elas foram acrescentadas a linha em negrito, cor vermelha, com o resultado da PGE para cada conjunto de dados.

Tabela 5.10 – Algoritmos de classificação para o conjunto de dados Wisconsin

Posição	Algoritmo	%-corr	Referência
1	NB + kernel est	97,5±1,8	WD, WEKA, 10X10CV
2	SVM (5xCV)	97,2	Bennet and Blue
3	kNN with DVDM distance	97,1	our (KG)
4	GM k-NN, k=3, raw, Manh	97,0±2,1	WD, 10X10CV
5	GM k-NN, k=opt, raw, Manh	97,0±1,7	WD, 10CV only
6	VSS, 8 it/2 neurons	96,9±1,8	WD/MK; 98.1% train
7	FSM-Feature Space Mapping	96,9±1,4	RA/WD, a=.99 Gaussian
8	Fisher linear discr. anal	96,8	Ster, Dobnikar
9	MLP+BP	96,7	Ster, Dobnikar
10	MLP+BP (Tooldiag)	96,6	Rafał Adamczak
11	LVQ	96,6	Ster, Dobnikar
12	kNN, Euclidean/Manhattan f.	96,6	Ster, Dobnikar
13	SNB, semi-naive Bayes (pairwise dependent)	96,6	Ster, Dobnikar
14	**PGE**	**96,43±2,88**	
15	SVM lin, opt C	96,4±1,2	WD-GM, 16 missing with -10
16	VSS, 8 it/1 neuron!	96,4±2,0	WD/MK, train 98.0%
17	GM IncNet	96,4±2,1	NJ/WD; FKF, max. 3 neurons
18	NB - naive Bayes (completly independent)	96,4	Ster, Dobnikar
19	SSV opt nodes, 3CV int	96,3±2,2	WD/GM; training 96.6±0.5
20	IB1	96,3±1,9	Zarndt
21	DB-CART (decision tree)	96,2	Shang, Breiman
22	GM SSV Tree, opt nodes BFS	96,0±2,9	WD/KG (beam search 94.0)
23	LDA - linear discriminant analysis	96	Ster, Dobnikar
24	OC1 DT (5xCV)	95,9	Bennet and Blue
25	RBF (Tooldiag)	95,9	Rafał Adamczak
26	GTO DT (5xCV)	95,7	Bennet and Blue
27	ASI - Assistant I tree	95,6	Ster, Dobnikar
28	MLP+BP (Weka)	95,4±0,2	TW/WD
29	OCN2	95,2±2,1	Zarndt
30	IB3	95,0±4,0	Zarndt
31	MML tree	94,8±1,8	Zarndt
32	ASR - Assistant R (RELIEF criterion) tree	94,7	Ster, Dobnikar
33	C4.5 tree	94,7±2,0	Zarndt
34	LFC, Lookahead Feature Constr binary tree	94,4	Ster, Dobnikar
35	CART tree	94,4±2,4	Zarndt
36	ID3	94,3±2,6	Zarndt
37	C4.5 (5xCV)	93,4	Bennet and Blue
38	C 4.5 rules	86,7±5,9	Zarndt
39	Default, majority	65,5	--
40	QDA - quadratic discr anal	34,5	Ster, Dobnikar

Tabela 5.11 – Algoritmos de classificação para o conjunto de dados Diabetes

Posição	Algoritmo	%-corr	Referência
1	Logdisc	77,7	Statlog
2	IncNet	77,6	Norbert Jankowski
3	DIPOL92	77,6	Statlog
4	Linear Discr. Anal.	77,5-77,2	Statlog; Ster & Dobnikar
5	SVM, linear, C=0.01	77,5±4,2	WD-GM, 10XCV averaged 10x
6	SVM, Gauss, C, sigma opt	77,4±4,3	WD-GM, 10XCV averaged 10x
7	PGE	76,95±6,00	
8	SMART	76,8	Statlog
9	GTO DT (5xCV)	76,8	Bennet and Blue
10	kNN, k=23, Manh, raw, W	76,7±4,0	WD-GM, feature weighting 3CV
11	kNN, k=1:25, Manh, raw	76,6±3,4	WD-GM, most cases k=23
12	ASI	76,6	Ster & Dobnikar
13	Fisher discr. analysis	76,5	Ster & Dobnikar
14	MLP+BP	76,4	Ster & Dobnikar
15	MLP+BP	75,8±6,2	Zarndt
16	LVQ	75,8	Ster & Dobnikar
17	LFC	75,8	Ster & Dobnikar
18	RBF	75,7	Statlog
19	NB	75,5-73,8	Ster & Dobnikar; Statlog
20	kNN, k=22, Manh	75,5	Karol Grudziński
21	MML	75,5±6,3	Zarndt
22	FSM stand. 5 feat.	75,4±4,9	WD, 10x10 test, CC>0.15
23	SNB	75,4	Ster & Dobnikar
24	BP	75,2	Statlog
25	SSV DT	75,0±3,6	WD-GM, SSV BS, node 5CV MC
26	kNN, k=18, Euclid, raw	74,8±4,8	WD-GM
27	CART DT	74,7±5,4	Zarndt
28	CART DT	74,5	Stalog
29	DB-CART	74,4	Shang & Breiman
30	ASR	74,3	Ster & Dobnikar
31	FSM standard	74,1±1,1	WD, 10x10 test
32	ODT, dyadic trees	74,0±2,3	Blanchard
33	Cluster means, 2 prototypes	73,7±3,7	MB
34	SSV DT	73,7±4,7	WD-GM, SSV BS, node 10CV strat
35	SFC, stacking filters	73,3±1,9	Porter
36	C4.5 DT	73	Stalog
37	C4.5 DT	72,7±6,6	Zarndt
38	Bayes	72,2±6,9	Zarndt
39	C4.5 (5xCV)	72	Bennet and Blue
40	CART	72,8	Ster & Dobnikar
41	Kohonen	72,7	Statlog
42	C4.5 DT	72,1±2,6	Blanchard (averaged over 100 runs)
43	kNN	71,9	Ster & Dobnikar
44	ID3	71,7±6,6	Zarndt
45	IB3	71,7±5,0	Zarndt
46	IB1	70,4±6,2	Zarndt
47	kNN, k=1, Euclides, raw	69,4±4,4	WD-GM
48	kNN	67,6	Statlog
49	C4.5 rules	67,0±2,9	Zarndt
50	OCN2	65,1±1,1	Zarndt
51	Default, majority	65,1	
52	QDA	59,5	Ster, Dobnikar

Tabela 5.12 – Algoritmos de classificação para o conjunto de dados Ionosfera

Posição	Algoritmo	%-corr	Referência
1	3-NN + simplex	98,7	Our own weighted kNN
2	VSS 2 epochs	96,7	MLP with numerical gradient
3	3-NN	96,7	KG, GM with or without weights
4	IB3	96,7	Aha, 5 errors on test
5	1-NN, Manhattan	96	GM kNN (our)
6	MLP+BP	96	Sigillito
7	SVM Gaussian	94,9±2,6	GM (our), defaults, similar for C=1-100
8	C4.5	94,9	Hamilton
9	3-NN Canberra	94,7	GM kNN (our)
10	RIAC	94,6	Hamilton
11	C4 (no windowing)	94	Aha
12	C4.5	93,7	Bennet and Blue
13	SVM	93,2	Bennet and Blue
14	Non-lin perceptron	92	Sigillito
15	FSM + rotation	92,8	our
16	1-NN, Euclidean	92,1	Aha, GM kNN (our)
17	DB-CART	91,3	Shang, Breiman
18	Linear perceptron	90,7	Sigillito
19	OC1 DT	89,5	Bennet and Blue
20	CART	88,9	Shang, Breiman
21	SVM linear	87,1±3,9	GM (our), defaults
22	PGE	86,9+5,21	
23	GTO DT	86	Bennet and Blue

A PGE apresentou competitividade frente aos outros algoritmos, nos conjuntos de dados Wisconsin e Diabetes. Muitos dos algoritmos que se posicionaram ligeiramente a frente da PGE (com diferença percentual média de acertos em relação à PGE inferior a 0,5%) nos dois conjuntos de dados citados, como o SVM (*Support Vector Machine*) e MLP+BP (*Multilayer Perceptron* com *Back Propagation*), são altamente não lineares, utilizando-se de funções de maior complexidade tais como as funções trigonométricas, permitindo pouca ou nenhuma interpretabilidade de resultados e da estrutura de seus modelos. Para SVM e MLP+BP, veja Hastie et al. (2011) e Bishop (1996), respectivamente.

Os modelos logit, utilizados pela PGE, embora não lineares, são lineares na estrutura de seus regressores, utilizando-se somente das operações de soma e multiplicação para combiná-los. A PGE se utiliza de ferramental menos abrangente que o SVM e o MLP+BP, por exemplo, mas consegue competir com algoritmos deste nível nos conjuntos de dados citados.

Em Ionosfera, o desempenho da PGE não foi competitivo. Ionosfera possui um atributo binário, uma das variáveis independentes que constitui o

conjunto de dados, que apresenta 89,17% de suas ocorrências com saída igual a 1 – o restante, 10,83% das ocorrências, possui saída igual à zero. Esse atributo pode se combinar com outros, via multiplicação, dando origem a regressores que possuam relação de dependência elevada entre si, fazendo com que o posto de $\boldsymbol{X}$ não seja cheio.

A variante do MN utilizada nesta dissertação é uma generalização do MN para solucionar o sistema de equações em (2.24), para maximização de $l(\boldsymbol{\beta})$, e não apresenta uma "proteção" contra $\boldsymbol{X}$ que não tenha posto cheio, como possui a decomposição QR para minimizar o SQR, que elimina colunas de $\boldsymbol{X}$ linearmente dependentes. Com isso, a PGE é obrigada a utilizar parâmetros de experimento tais que não criem indivíduos muito grandes, para que se evite obter indivíduos com $\boldsymbol{X}$ que não tenha posto cheio, para conjuntos de dados em tarefas de classificação nas condições citadas. Dessa forma, indivíduos com acurácia maior deixam de ser gerados por uma limitação da PGE, sujeita a conjuntos de dados altamente desbalanceados, como é o caso de Ionosfera.

6
Conclusão

Pode-se dizer que a PGE cumpriu satisfatoriamente com seu objetivo principal, ao propor modelos parcimoniosos de elevada acurácia para os conjuntos de dados propostos – os valores nominais das métricas de acurácia, dispostas no capítulo anterior, evidenciam a assertiva – e promoveu modelos competitivos frente a outros algoritmos de regressão e classificação – avalia-se a assertiva pela análise do desempenho comparativo aos *benchmarks*, principalmente em Concreto, Casas, Iates, Wisconsin e Diabetes.

Observou-se que a PGE não forneceu modelos de elevada acurácia em todos os conjuntos de dados: por exemplo, o $\overline{R}^2$ médio dos melhores indivíduos gerados para Proteínas, no conjunto de teste, foi abaixo de 0,30. Além disso, observou-se que a PGE não foi competitiva em todos os casos: por exemplo, seu desempenho em Ionosfera foi tal que gerou ao algoritmo uma das últimas colocações entre os algoritmos classificadores. Portanto, há benefícios e limitações relativas ao seu uso como processo gerador de modelos para regressão e classificação.

Sob o domínio da computação evolucionária e da PG, destacam-se como benefícios da PGE:

(1) O auxílio na identificação de *introns* através da significância estatística. Miller & Smith (2006) definem *introns* como partes estranhas do código que não contribuem à acurácia do indivíduo. Ao se realizar TH em $\boldsymbol{\beta}$, retira-se do cômputo da acurácia os regressores estatisticamente insignificantes – que são os *introns*, segundo a definição de Miller & Smith (2006).

(2) O combate ao *bloat* através da significância estatística. Luke & Panait (2006) definem *bloat* como o crescimento ilimitado e sem controle de indivíduos em uma população, geralmente não ocasionando melhorias na acurácia. Embora se permita que *introns* permaneçam na estrutura de um indivíduo, potencializando o *bloat*, a PGE o combate quando define que somente o regressor estatisticamente significante contribui para a acurácia. Para a PGE, o crescimento generalizado de indivíduos pode até não ser visto como um problema, pois ela se beneficia de

potenciais regressores estatisticamente significantes que venham a surgir a partir deste crescimento.

(3) A geração de modelos lineares, para regressão, e de não lineares, para classificação, com uma porção linear que permite análise simples da estrutura do modelo e regressores que o compõem.

(4) A alternativa à utilização de constantes efêmeras, através da estimação de $\widehat{\boldsymbol{\beta}}$ por MQO ou MV. O uso de $\widehat{\boldsymbol{\beta}}$ em substituição às constantes efêmeras em um modelo potencializa a contribuição, por meio da acurácia, deste modelo à tarefa em questão, pois a estimação de $\widehat{\boldsymbol{\beta}}$ é um processo de otimização, que permite ao modelo estar em suas melhores condições (em função do critério que se utiliza para otimização de $\widehat{\boldsymbol{\beta}}$) de ser aplicado à tarefa.

Sob o domínio da computação evolucionária e da PG, destacam-se como limitações da PGE:

(1) A dupla estimação de $\widehat{\boldsymbol{\beta}}$ para cada indivíduo, necessária à identificação de regressores estatisticamente significantes e do cômputo da acurácia.

(2) A restrição relacionada à forma dos modelos $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ e $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ pode inibir a geração de outros modelos, menos restritivos em sua forma, mas que poderiam apresentar melhor acurácia. As formas $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ e $P_t = \Lambda(\boldsymbol{X}_t\boldsymbol{\beta})$ também implicam, em primeira instância, em se utilizar somente as funções soma e multiplicação no conjunto de funções.

Sob o domínio da econometria, destacam-se como benefícios da PGE:

(1) A geração de modelos com foco em elevada acurácia. Tanto os modelos gerados por PGE quanto seus *benchmarks* estão altamente comprometidos com a busca pela melhor assertividade que se possa fornecer à tarefa do conjunto de dados em questão.

(2) Intuitivamente, um econometrista ou outro usuário da PGE poderia não pensar em uma determinada combinação de variáveis de Ω como um regressor estatisticamente significante. Supondo que a PGE gere um modelo, ao fim da evolução, que contenha este regressor, a situação pode fazer com que o econometrista ou outro usuário reflita sobre a racionalidade – econômica ou física, por exemplo – desta combinação, permitindo que se possa relacionar a acurácia fornecida pelo modelo, em função de seus regressores, com causalidade.

Sob o domínio da econometria, destacam-se como limitações da PGE:

(1) Possibilidade de pouca interpretabilidade de $\widehat{\boldsymbol{\beta}}$. Por exemplo, o par amostral do modelo populacional a seguir foi eleito o melhor indivíduo para um dos experimentos com o conjunto de dados Casas.

$$\begin{aligned} y = {} & \beta_1 x_1 + \beta_2 x_5 + \beta_3 x_6 + \beta_4 x_7 + \beta_5 x_9 + \beta_6 x_{11} + \beta_7 x_6 x_9 + \beta_8 x_6 x_{13} \\ & + \beta_9 x_9 x_{13} + \beta_{10} x_{13}^2 + \beta_{11} x_8 x_9 x_{11}^2 + \beta_{12} x_3 x_9 x_{10} \\ & + \beta_{13} x_3 x_{10} x_{13} + \beta_{14} x_8 x_9 x_{11} + \beta_{15} x_9 x_{10} x_{11} \\ & + \beta_{16} x_{10} x_{11} x_{13} + u \end{aligned} \tag{6.1}$$

É possível que haja interesse na interpretação do coeficiente do regressor $x_{10}x_{11}x_{13}$, constituído pelas variáveis x_{10}, x_{11} e x_{13}, de Ω. Embora seja estatisticamente significante, é possível que $\hat{\beta}_{16}$ seja não interpretável dada a estrutura de regressores do par amostral de (6.1).

Para o mesmo experimento citado acima, que gerou o par amostral do indivíduo de (6.1), gerou-se o par amostral do indivíduo em (6.2).

$$y = \beta_1 x_1 + \beta_2 x_4 + \beta_3 x_5 + \beta_4 x_6 + \beta_5 x_{11} + \beta_6 x_{12} + u \tag{6.2}$$

Ao passo que o par amostral de (6.1) apresentou $\overline{\mathrm{R}}^2 = 0{,}8615$ no conjunto de treino, com cardinalidade de X igual a 16, o par amostral de (6.2) apresentou $\overline{\mathrm{R}}^2 = 0{,}5791$, com cardinalidade de X igual a 6. É possível que um usuário da PGE opte pela utilização de (6.2) em detrimento de (6.1), por (6.2) apresentar uma estrutura menor com potencial de interpretabilidade de coeficientes maior. Portanto, embora a PGE possa fornecer modelos com pouca interpretabilidade de $\widehat{\boldsymbol{\beta}}$, ela também tem o potencial de gerar modelos com alta capacidade de interpretabilidade, cabendo ao usuário a decisão de qual modelo utilizar.

(2) Dificuldade em obter bom desempenho em conjuntos de dados que sejam altamente correlacionados ou que apresentem atributos não informativos, pelo fato da variante do MN, a partir de sua generalização para o sistema de equações em (2.24), para maximização de $l(\boldsymbol{\beta})$, não apresentar uma proteção contra $\boldsymbol{X}$ que não tenha posto cheio, como possui a decomposição QR para minimizar o SQR, que elimina colunas de $\boldsymbol{X}$ linearmente dependentes. Com isso, a

PGE é obrigada a utilizar parâmetros de experimento tais que não criem indivíduos muito grandes, para que se evite obter indivíduos com $\boldsymbol{X}$ que não tenha posto cheio. Tal limitação impossibilita a geração de indivíduos com acurácia mais elevada em conjuntos de dados dentro das condições citadas.

6.1
Desenvolvimentos Futuros

Desenvolvimentos futuros serão baseados nas seguintes ações:

(1) Inserção de TH para $\boldsymbol{\beta}$ com a variância de White. Embora a tabela 4.1, que evidencia os resultados para TH em regressores com variâncias distintas para 5.000 indivíduos de cada conjunto de dados, ter apresentado similaridade alta entre os resultados, nada se pode afirmar com relação a como se comportará a variância em novos e distintos conjuntos de dados, sendo adequado que se utilize a variância de White.

(2) Expansão da PGE para conjuntos de dados em séries de tempo.

(3) Possibilidade de evoluir paralelamente à PGE um algoritmo genético que selecione somente variáveis independentes simples de Ω. Por exemplo, em (6.1), há poucas variáveis independentes simples de Ω como regressores. Como há a tendência, em um experimento de PGE, que sejam sugeridos mais regressores combinados do que regressores simples aos modelos, o algoritmo genético seria uma ferramenta adequada para sugerir regressores simples, evoluindo paralelamente à PGE, calculando-se a acurácia dos modelos utilizando tanto os regressores sugeridos pela PGE quanto pelo algoritmo genético.

7

Referências Bibliográficas

ALTMAN, M.; GILL, J.; McDONALD, M.P. **Numerical Issues in Statistical Computing for the Social Scientist**. 1. ed. New Jersey: Wiley-Interscience, 2003.

AMEMIYA, T. **Advanced Econometrics**. 1. ed. [S.l.]: Harvard University Press, 1985.

ARMSTRONG, J. S.; FILDES, R. **Correspondence: On the selection of error measures for comparisons among forecasting methods**. *Journal of Forecasting*, v. 14, p. 67-72, 1995.

ASHLAGI, I. et al. **Monotonicity and Implementability**, *Econometrica*, v. 78, n. 5, p. 1749-1772, set. 2010.

BISHOP, C.M. **Neural Networks for Pattern Recognition (Advanced Texts in Econometrics)**. 1. ed. [S.l.]: Oxford University Press, 1996.

BJÖRCK, Â. **Solving Linear Least Squares Problems By Gram-Schmidt Orthogonalization**, *BIT*, v. 7, p. 1-21, 1967.

BORWEIN, J.M.; LEWIS, A.S. **Convex Analysis and Nonlinear Optimization: Theory and Examples**. 2. ed. [S.l.]: Springer, 2005.

BURDEN, R.L.; FAIRES, J.D. **Numerical Analysis**. 9. ed. [S.l.]: *Brooks Cole*, 2011.

BUSINGER, P.; GOLUB, G.H. **Linear Least Squares Solutions by Householder Transformations**, *Numerische Mathematik*, v. 7, p. 269-276, 1965.

CASELLA, G.; BERGER, R.L. **Inferência Estatística**. Tradução de Solange A. Visconte. 2. ed. [S.l.]: Cengage Learning, 2011.

CHAMBERS, J.M. **Computational Methods for Data Analysis (Probability & Mathematical Statistics)**. 1. ed. New York: John Wiley & Sons, 1977.

DAVIDSON, R.; MacKINNON, J.G. **Estimation and Inference in Econometrics**. 1. ed. [S.l.]: Oxford University Press, 1993.

DAVIDSON, R.; MacKINNON, J.G. **Econometric Theory and Methods**. 1. ed. [S.l.]: Oxford University Press, 2003.

DAVIDSON, J. W. et al. **Method for the Identification of Explicit Polynomial Formulae for the Friction in Turbulent Pipe Flow**. *Journal of Hydroinformatics*, v. 1, n. 2, p. 115-126, 1999.

DAVIS, L. **Adapting operator probabilities in genetic algorithms**. In: The Third International Conference on Genetic Algorithms, 3., 1989, [S.l.]. *Proceedings of the Third International Conference on Genetic Algorithms*, J. David Schaffer (Ed.) (Morgan Kaufmann Publishers, San Mateo), p. 61–69.

DENISON, D.G.T. et al. **Bayesian Methods for Nonlinear Classification and Regression**. 1. ed. [S.l.]: Wiley, 2002.

DEUFLHARD, P. **A Modified Newton Method for The Solution of Ill-Conditioned Systems of Nonlinear Equations with Application to Multiple Shooting**, *Numer Math*, v. 22, p. 289-315, 1974.

DOMINGOS, P. **The Role of Occam's Razor in Knowledge Discovery**. *Data Mining and Knowledge Discovery*, v. 3, p. 409-425, 1999.

DRAPER, N.R.; SMITH, H. **Applied Regression Analysis**. 3. ed. [S.l.]: Wiley-Interscience, 1998.

EICKER, F. **Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions**, *The Annals of Mathematical Statistics*, v. 34, p. 447-456, 1963.

EICKER, F. **Limit Theorems for Regressions with Unequal and Dependent Errors**. In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1., 1967, Berkeley. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, p. 59–82.

FERNANDES, C.A.C. **Regressão Logística para Dados Binários**. 1. ed. Rio de Janeiro, RJ: [s.n.], 2009.

GANDOMI, A.H.; ALAVI, A.H. **Multi-stage genetic programming: a new strategy to nonlinear system modeling**, *Information Sciences*, v. 181, n. 23, p. 5227-5239, 2011.

GIUSTOLISI, O.; SAVIC, D.A. **A symbolic data-driven technique based on evolutionary polynomial regression**, *Journal of Hydroinformatics*, v. 8, n. 3, p. 207-222, 2006.

GIUSTOLISI, O. **Using genetic programming to determine Chèzy resistance coefficient in corrugated channels**, *Journal of Hydroinformatics*, v. 3, n. 6, p. 157-173, 2004.

GOLDBERGER, A.S. **A Course in Econometrics**. 1. ed. [S.l.]: Harvard University Press, 1991.

GRAVES, L.M. **Riemann Integration and Taylor's Theorem in General Analysis**. *Transactions of the American Mathematical Society*, v. 29, p. 163-177, 1927.

GREENE, W.H. *Econometric Analysis*. 7. ed. [S.l.]: Prentice Hall, 2011.

GUJARATI, D.; PORTER, D. *Basic Econometrics*. 5. ed. [S.l.]: McGraw-Hill/Irwin, 2008.

HAAVELMO, T. **The Probability Approach in Econometrics**, *Econometrica*, v. 12, p. 1-118, 1944. Suplemento.

HANSON, R.J.; LAWSON, C.L. **Extensions and Applications of the Householder Algorithm for Solving Linear Least Squares Problems**, *Mathematics of Computation*, v. 23, n. 108, p. 787-812, out. 1969.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. [S.l.]: Springer, 2011.

HEDRICK, J.K.; GIRARD, A. **Control of Nonlinear Dynamic Systems: Theory and Applications**. 1. ed. [S.l.]: [s.n.], 2010.

HINES, W.W. et al. **Probability and Statistics in Engineering**. 4. ed. [S.l.]: Wiley, 2003.

HOLLAND, J.H. **Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence** (Complex Adaptive Systems). 2. ed. [S.l.]: MIT Press, 1992.

HUBER, P.J. **The behaviour of maximum likelihood estimates under non-standard conditions**. In: Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1., 1967, Berkeley. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, p. 221–233.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: Fourteenth International Joint Conference on Artificial Intelligence, 1., 1995, San Francisco, CA. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, San Francisco: [s.n.], p. 1137–1143.

KOLMOGOROV, A.N. **Foundations of the Theory of Probability**. 2. ed. [S.l.]: Chelsea Pub Co, 1960.

KOOPMANS, T. C. **The Equivalence of Maximum Likelihood and Least Squares Estimates of Regression Coefficients**. In: *Statistical Inference in Dynamic Economic Models*, 1. ed. New York: Cowles Foundation for Research in Economics, 1950. Cap. 7.

KOZA, J.R. **Genetic Programming**: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems). 1. ed. [S.l.]: The MIT Press, 1992.

LICHMAN, M. **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

LING, F. **Givens Rotation Based Least Squares Lattice and Related Algorithms**. *IEEE Transactions on Signal Processing*, v. 39, n. 7, p. 1541-1551, jul. 1991.

LING, F. et al. **A Recursive Modified Gram-Schmidt Algorithm For Least-Squares Estimation**. *IEEE Transactions on Acoustics, Speech, And Signal Processing*, v. 34, n. 4, p. 829-836, ago. 1986.

LJUNG, L. **System Identification**: Theory for the User. 2. ed. [S.l.]: Prentice Hall, 1999.

LUKE, S.; PANAIT, L. **Lexicographic parsimony pressure**. In: GECCO-2002, 1., 2002, San Francisco, CA. *Proceedings of GECCO-2002*, San Francisco: [s.n.], p. 829–836.

LUKE, S.; PANAIT, L. **A comparison of bloat control methods for genetic programming**. *Evolutionary Computation*, v. 14, n. 3, p. 309-344, 2006.

MAINDONALD, J.H. **Statistical Computation**. 1. ed. New York: Wiley, 1984.

MARDIA, K.V. et al. **Multivariate Analysis**. 1. ed. [S.l.]: Academic Press, 1980.

MARSAGLIA, G.; MARSAGLIA, J.C.W. **A New Derivation of Stirling's Approximation to n!**. *The American Mathematical Monthly*, v. 97, n. 9, p. 826-829, nov. 1990.

MILLER, J.F.; SMITH, S.L. **Redundancy and Computational Efficiency in Cartesian Genetic Programming**. *IEEE Transactions on Evolutionary Computation*, v. 10, n. 2, p. 167-174, abril 2006.

MURRAY, W. **Newton-type Methods**. 1. ed. Stanford, CA: [s.n.], 2010.

MYUNG, I.J.; PITT, M.A. **Applying Occam's razor in modeling cognition: A Bayesian approach**. *Psychonomic Bulletin & Review*, v. 4, n. 1, p. 79-95, 1997.

POLI, R. et al. **A Field Guide to Genetic Programming**. 1. ed. [S.l.]: Lulu Enterprises, UK Ltd, 2008.

POOLE, D. **Linear Algebra: A Modern Introduction**. 3. ed. [S.l.]: Cengage Learning, 2010.

PRATT, J.W. **Concavity of the log likelihood**. *Journal of the American Statistical Association*, v. 76, p. 103-106, 1981.

SEARSON, D.P.; LEAHY, D.E.; WILLIS, M.J. **GPTIPS: an open source genetic programming toolbox for multigene symbolic regression**. In: The International MultiConference of Engineers and Computer Scientists 2010 (IMECS 2010), 1., 2010, Hong Kong. *Proceedings of IMECS 2010*, Hong Kong: [s.n.].

SILVA, S. **GPLAB: A Genetic Programming Toolbox for MATLAB**. 3. ed. [S.l.]: [s.n.], 2007.

SPANOS, A. **Probability Theory and Statistical Inference: Econometric Modeling with Observational Data**. 1. ed. [S.l.]: Cambridge University Press, 1999.

TINTNER, G. **Methodology of Mathematical Economics and Econometrics**. 1. ed. [S.l.]: University of Chicago Press, 1968.

UNIVERSIDADE DE NICOLAU COPÉRNICO. Laboratório de Inteligência Computacional do Departamento de Informática. *Datasets used for classification: comparison of results*. [S.l.], Novembro de 2014.

WANG, Z.; BOVIK, A.C. **Mean squared error: Love it or leave it? – A new look at signal fidelity measures**. *IEEE Signal Processing Magazine*, v. 26, n. 1, p. 98-117, jan. 2009.

WHITE, H. **A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity**. *Econometrica*, v. 48, p. 817-838, 1980.

WILLCOX, W.F. **The Founder of Statistics**. *Review of the International Statistical Institute*, v. 5, n. 4, p. 321-328, jan. 1938.

WOOLDRIDGE, J.M. **Econometric Analysis of Cross Section and Panel Data**. 1. ed. [S.l.]: *The MIT Press*, 2001.

WOOLDRIDGE, J.M. **Introdução à Econometria: Uma Abordagem Moderna**. Tradução de Rogério César de Souza e José Antônio Ferreira. 1. ed. [S.l.]: Thomson Heinle, 2006.

WOOLDRIDGE, J.M. **Introductory Econometrics: A Modern Approach**. 4. ed. [S.l.]: Cengage Learning, 2008.