5
O Modelo STR-Tree

5.1
Introdução

Aqui introduz-se um modelo de regressão não-linear, estruturado por uma árvore de decisão binária. Este modelo será denominado STR-Tree(Smooth Transition Regression Tree-structured). A principal idéia do modelo STR-Tree é aproveitar vantagens da metodologia CART apresentada no Capítulo 3, mas introduzir elementos que tornem possível o uso de ferramentas inferenciais. A intenção é manter a interpretabilidade dos modelos estruturados por árvores, mas analisá-los como uma classe de modelos não-lineares paramétricos.

A forma funcional altamente descontínua dos modelos ajustados pelo algoritmo CART e a estratégia que força a diminuição da soma dos erros quadráticos através de sucessivas divisões da amostra, impossibilitam a implementação de testes de significância estatística e a utilização de elementos da inferência clássica.

A idéia a ser utilizada aqui é a mesma em [15] e [83]; a substituição de divisões abruptas utilizadas na modelagem CART por divisões suaves.

Os desenvolvimentos a serem apresentados neste capítulo seguem toda a estrutura apresentada no Capítulo 3. Considere, por exemplo, a árvore mais simples com dois nós terminais gerada em (3-4). Ao substituir a função indicadora $I(\cdot)$ em (3-4) pela função logística definida como:

$$G(\mathbf{x}_t; s_0, \gamma_0, c_0) = \frac{e^{-\gamma_0\left(x_{s_0 t} - c_0\right)}}{1 + e^{-\gamma_0\left(x_{s_0 t} - c_0\right)}}, \tag{5-1}$$

obtém-se

$$y_t = \beta_1 G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \beta_2 \left[1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)\right] + \varepsilon_t, \tag{5-2}$$

onde agora há a presença de um parâmetro adicional $\gamma_0 > 0$ que controla a suavidade da função logística.

Esta mudança acarreta uma importante diferença em relação à abordagem CART; a divisão do nó raiz não irá gerar dois subconjuntos exclusivos, mas irá criar dois conjuntos nebulosos [98] aos quais todas as observações irão pertencer, porém com diferentes graus de pertinência.

Repare que o tipo de partição feita pelo CART torna-se um caso particular quando o parâmetro de suavidade aproxima-se de infinito. Por outro lado, quando este parâmetro se aproxima de zero, gera-se a situação mais nebulosa na qual não haverá ganho em produzir a divisão do nó principal. O parâmetro c_0 será chamado aqui de *parâmetro de locação*.

Assumindo que o erro é uma variável aleatória com distribuição de probabilidade conhecida, de (5-1) e (5-2) torna-se possível, sem perder a flexibilidade da abordagem CART, interpretar a o modelo resultante como um caso particular dos modelos STR discutidos em [41] e [86] [1].

5.2 Construção do Modelo

O processo de construção do modelo STR-Tree é uma adaptação do ciclo de modelagem descrito em [85, 86]. Como mencionado na seção anterior, o objetivo é obter uma estratégia coerente para o crescimento da árvore utilizando inferência estatística. A arquitetura do modelo deve ser determinada a partir dos dados no estágio que aqui será chamado de *especificação* do modelo que envolve duas decisões: a seleção de um nó a ser dividido e o índice da variável de divisão a ser selecionado dentro do conjunto $\mathbb{S} = \{1, 2, \ldots, m\}$.

O estágio de especificação será conduzido por uma seqüência de testes do tipo Multiplicadores de Lagrange (ML) seguindo as idéias originalmente apresentadas em [64]. Uma abordagem alternativa baseada em validação cruzada com 10-dobras também é possível. Entretanto, o gasto computacional envolvido é dramaticamente alto. Maiores detalhes são apresentados na Seção 5.5.

O estágio de especificação requer também *estimação* dos parâmetros do modelo. Após a especificação, seguida da estimação dos parâmetros, o próximo passo no ciclo de modelagem é a *avaliação* do modelo estimado. Em geral, modelos estruturados por árvores são avaliados pelo seu desempenho (habilidade preditiva) em dados fora-da-amostra (*out-of-sample*). Aqui, será seguida a literatura e o modelo STR-Tree será avaliado da mesma forma. A construção de testes para falhas de especificação no mesmo espírito dos apresentados em [35] é possível, mas não serão abordados nesta tese.

[1] O modelo STR-Tree pode também ser visto como um caso particular dos modelos MRSTAR (Multiple-Regime Smooth Transition Autoregressive) apresentados em [103].

Seguindo o pr"específico-para-geral", o ciclo de modelagem incia-se dentro do nó raiz (profundidade 0) e os passos gerais são:

1. Especificação do modelo selecionando através do teste tipo ML, dentro da profundidade d, um nó a ser dividido (caso $d > 0$) e uma variável de divisão.

2. Estimação dos parâmetros da função logística e das constantes dentro das folhas.

3. Avaliação do modelo estimado checando se é necessário:

   (a) Alterar o nó a ser dividido.

   (b) Alterar a variável de divisão.

   (c) Remover a divisão.

4. Utilizar o modelo final com o propósito de fazer predições ou simplesmente para uso descritivo.

A Figura 5.1 ilustra o ciclo de modelagem. Este começa da raiz (nó 0) aplicando um teste de hipóteses do modelo constante global contra a alternativa do mais simples modelo STR-Tree que contém apenas 2 nós terminais.

Como a seleção da arquitetura da árvore está ligada diretamente à estimação de parâmetros, as atenções serão voltadas agora para este tópico.

5.3 Estimação de Parâmetros

Considere um modelo STR-Tree em uma árvore completamente crescida. Isto significa dizer que na profundidade d existem $K = 2^d$ nós terminais (folhas), que é o número máximo, e $N = \sum_{i=1}^{d} 2^i$ nós geradores. A formulação matemática deste modelo é:

$$y_t = H\left(\mathbf{x}_t; \boldsymbol{\psi}\right) = \sum_{k=1}^{K} \beta_{K+k-2} B_k(\mathbf{x}_t; \boldsymbol{\theta}_k) + \varepsilon_t, \tag{5-3}$$

onde $B_k(\mathbf{x}_t; \boldsymbol{\theta}_k)$, $k = 1, \ldots, K$, é definido pelo produto de funções logísticas.

O vetor de parâmetros $\boldsymbol{\psi} = (\beta_{K-1}, \ldots, \beta_{2K-2}, \boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K)'$ tem $r = K + 2N$ elementos.

Como um exemplo desta situação, a arquitetura da árvore na Figura 3.6 está associada a um modelo STR-Tree que tem profundidade $d = 2$, $K = 4$, $N = 3$, e as funções:

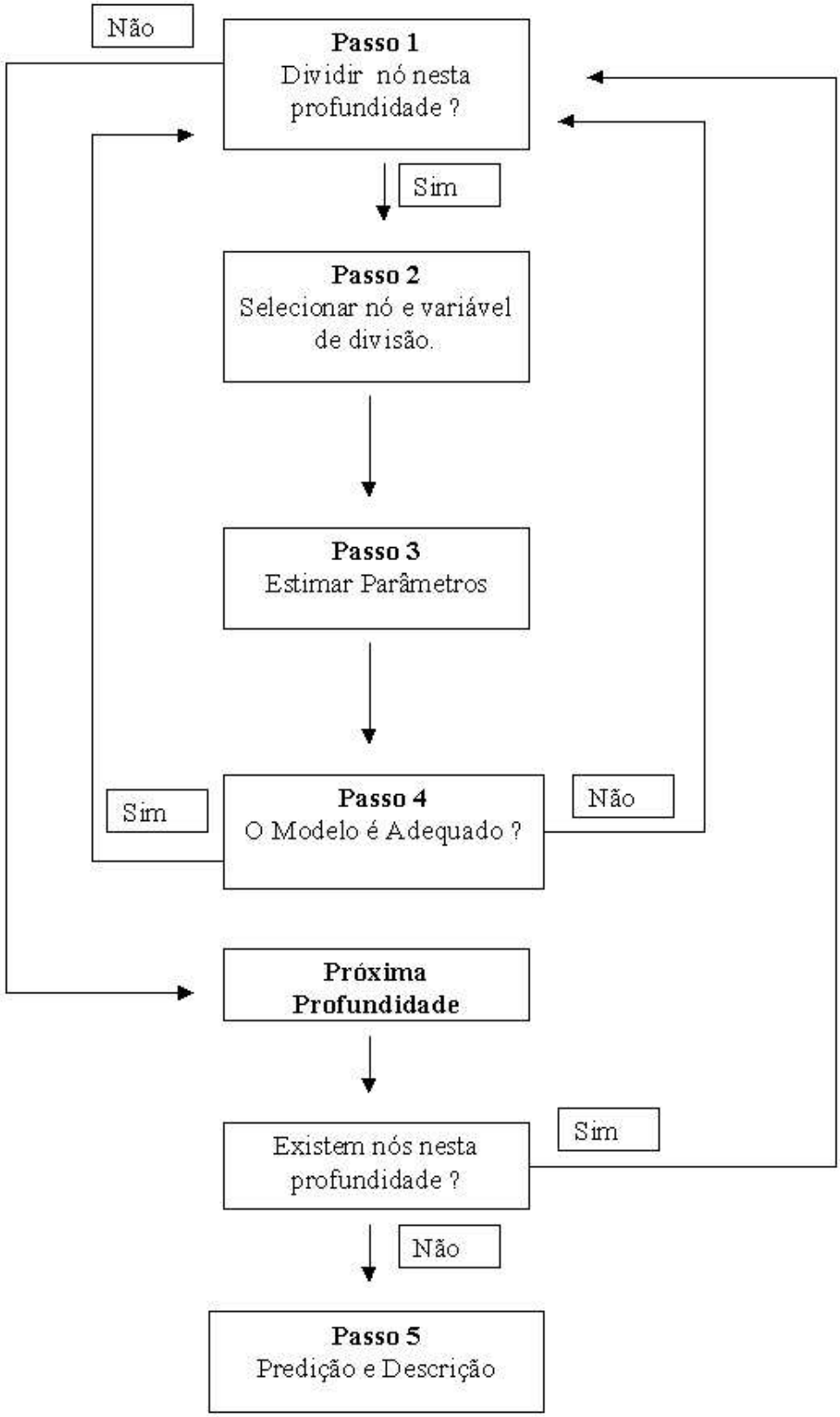

Figura 5.1: Ciclo de Modelagem do modelo STR-Tree

$$B_1(\mathbf{x}_t;\boldsymbol{\theta}_1) = G(\mathbf{x}_t;s_0,\gamma_0,c_0)\,G(\mathbf{x}_t;s_1,\gamma_1,c_1);$$
$$B_2(\mathbf{x}_t;\boldsymbol{\theta}_2) = G(\mathbf{x}_t;s_0,\gamma_0,c_0)\,[1-G(\mathbf{x}_t;s_1,\gamma_1,c_1)];$$
$$B_3(\mathbf{x}_t;\boldsymbol{\theta}_3) = [1-G(\mathbf{x}_t;s_0,\gamma_0,c_0)]\,G(\mathbf{x}_t;s_2,\gamma_2,c_2) \text{ e}$$
$$B_4(\mathbf{x}_t;\boldsymbol{\theta}_4) = [1-G(\mathbf{x}_t;s_0,\gamma_0,c_0)]\,[1-G(\mathbf{x}_t;s_2,\gamma_2,c_2)].$$

Com esta arquitetura, o número total de parâmetros a ser estimado é 10 e existem 3 variáveis de transição para serem especificadas

Principais Hipóteses

Neste ponto, deve ser feito o seguinte conjunto de hipóteses.

Hipótese 1 *A seqüência $\{\mathbf{x}_t\}_{t=1}^{T}$ é formada por vetores aleatórios independentes e identicamente distribuídos e tem uma distribuição conjunta comum $\mathcal{D}$ em $\boldsymbol{\Delta}$, um espaço Euclideano, com densidade Radon-Nikodým mensurável.*

Hipótese 2 *A seqüência $\{\varepsilon_t\}_{t=1}^{T}$ é formada por variáveis aleatórias independentes e normalmente distribuídas (NID) com média zero e variância $\sigma^2 < \infty$, $\varepsilon_t \sim NID(0,\sigma^2)$.*

Hipótese 3 *O vetor $\boldsymbol{\psi}^*$, de dimensão $r \times 1$, que contém os verdadeiros valores dos parâmetros é um ponto interior de um espaço paramétrico compacto $\boldsymbol{\Psi}$ contido no $\mathbb{R}^r$, o espaço euclideano r-dimensional.*

Hipótese 4 *Os parâmetros γ_i, $i = 1,\ldots,N$, são números reais estritamente positivos onde N é o número de nós geradores. Além do mais, se para dois nós geradores adjacentes nas posições $2j+1$ e $2j+2$, $x_{s_{2j+1}t} = x_{s_{2j+2}t}$, então $c_{s_{2j+1}} < c_{s_{2j+2}}$.*

A Hipótese 1 simplesmente estabelece que o dados, sejam de painéis ou observações de uma série temporal, são IID. Embora a Hipótese 2 seja bem restritiva a primeira vista, o modelo (5-3) é ainda assim bastante flexível. Além do mais, esta suposição possibilita a utilização do princípio da máxima verossimilhança, que será equivalente ao de mínimos quadrados não-lineares. No caso de erros não gaussianos, a Hipótese 2 pode ser substituída por algumas condições nos momentos e uma pseudo-verossimilhança pode ser utilizada então. A principal diferença neste caso estará relacionada com o cálculo da matriz de covariância das estimativas dos parâmetros. Adicionalmente, uma versão robusta dos testes a serem apresentados pode ser construída no mesmo espírito de

[70], usando resultados desenvolvidos em [95]. A Hipótese 3 é padrão, enquanto a Hipótese 4 apenas garante que o modelo STR-Tree é identificável.

Como discutido préviamente, os parâmetros do modelo STR-Tree são estimados por Máxima Verossimilhança (MV) fazendo uso das suposições feitas sobre ε_t. O uso da máxima verossimilhança torna possível ter noção da incerteza sobre as estimativas dos parâmetros através do desvio padrão das estimativas. O modelo STR-Tree é similar a muitos modelos lineares e não-lineares no sentido de que a matriz de informação da função de log-verossimilhança é bloco-diagonal, de modo que a verossimilhança pode ser concentrada e primeiramente estimados os parâmetros da média condicional. A Máxima Verossimilhança Condicional é desta forma equivalente a Mínimos Quadrados Não-Lineares (MQNL).

O estimador de Mínimos Quadrados Não-Lineares do vetor de parâmetros é obtido por:

$$\widehat{\boldsymbol{\psi}} = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}}\, T^{-1}Q_T(\boldsymbol{\psi}) = \underset{\boldsymbol{\psi}\in\boldsymbol{\Psi}}{\operatorname{argmin}}\, T^{-1}\sum_{t=1}^{T} q_t(\boldsymbol{\psi}), \tag{5-4}$$

onde $q_t(\boldsymbol{\psi}) = [y_t - H(\mathbf{x}_t;\boldsymbol{\psi})]^2$.

Na seqüência, discute-se a existência, consistência e normalidade assintótica dos estimadores de MQNL definidos em (5-4).

Existência

A prova da existência dos estimadores de MQNL é baseada no Lema 2 apresentado em [58], que estabelece que sob certas condições de continuidade e mensurabilidade na função dos erros quadráticos médios (EQM), o estimador MQNL como mostrado em (5-4) existe.

Teorema 5.1 coloca as condições necessárias para a existência dos estimadores MQNL

Teorema 5.1 *O estimador MQNL existe se o modelo STR-Tree satisfaz as seguintes condições:*

1. *Para cada $\mathbf{x}_t \in \mathbb{X} \subseteq \mathbb{R}^m$, a função $H_{\mathbf{x}}(\boldsymbol{\psi}) = H(\mathbf{x}_t;\boldsymbol{\psi})$ é contínua em um subconjunto compacto Ψ do espaço euclideano.*

2. *Para cada $\boldsymbol{\psi} \in \Psi \subseteq \mathbb{R}^r$, a função $H_{\boldsymbol{\psi}}(\mathbb{X}) = H(\mathbf{x}_t;\boldsymbol{\psi})$ é mensurável no espaço $\mathbb{X}$.*

3. *ε_t são erros independentes e identicamente distribuídos com média zero e variância σ^2.*

Consistência

A consistência dos estimadores MQNL foi rigorosamente provada em [58] e [65]. Aqui seguem os resultados apresentados em [4] e colocado o seguinte teorema que garante condições sob as quais o estimador MQNL definido em 5-4 é fortemente consistente.

Teorema 5.2 *Sob as Hipóteses 1–5 o estimador MQNL $\widehat{\psi}$ é fortemente consistente para ψ^*, isto é ,$\widehat{\psi}$ converge quase certamente para ψ^*.*

Normalidade Assintótica

A normalidade assintótica dos estimadores MQNL foi cuidadosamente provada em [58]. Aqui segue-se o desenvolvimento de seus resultados e os desenvolvimentos em [4] que coloca o seguinte teorema.

Teorema 5.3 *Sob as Hipóteses 1–5*

$$T^{1/2}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}^*) \xrightarrow{d} N\left(\mathbf{0}, -\underset{T\to\infty}{plim}\ \mathbf{A}(\boldsymbol{\psi}^*)^{-1}\right), \tag{5-5}$$

onde $\mathbf{A}(\boldsymbol{\psi}^*) = \frac{1}{\sigma^2}\frac{\partial^2 Q_T(\boldsymbol{\psi}^*)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}$,$\xrightarrow{d}$ indica convergência em distribuição e plim denota convergência em probabilidade.

Comentário 1 *É direta a extensão dos teoremas acima para o caso de observações que não sejam IID e para modelos não corretamente especificados. Os resultados em [93], [94], e [96] podem ser aplicados.*

Mínimos Quadrados Concentrados

Condicionalmente ao conhecimento dos parâmetro $\boldsymbol{\theta}_k$ em (5-3), $k = 1, \ldots, K$, o modelo (5-3) é simplesmente uma regressão linear e o vetor de parâmetros $\boldsymbol{\beta} = (\beta_{K-1}, \ldots, \beta_{2K-2})'$ pode ser estimado por Mínimos Quadrados Ordinários (MQO) como:

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{B}(\boldsymbol{\theta})'\mathbf{B}(\boldsymbol{\theta})\right]^{-1}\mathbf{B}(\boldsymbol{\theta})'\mathbf{y}, \tag{5-6}$$

onde $\mathbf{y} = (y_1, \ldots, y_T)'$, $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_K)'$,
e

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{pmatrix} B_1(\mathbf{x}_1;\boldsymbol{\theta}_1) & \cdots & B_K(\mathbf{x}_1;\boldsymbol{\theta}_K) \\ \vdots & \ddots & \vdots \\ B_1(\mathbf{x}_T;\boldsymbol{\theta}_1) & \cdots & B_K(\mathbf{x}_T;\boldsymbol{\theta}_K) \end{pmatrix}.$$

Os parâmetros $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$, são estimados condicionalmente a $\boldsymbol{\beta}$ através da aplicação do algoritmo de Levenberg-Marquadt, ou de outro algoritmo de otimização não-convexa como o BFGS, completando desta forma a $i-$ésima iteração.

Como o algoritmo de MQNL é muito sensível à escolha dos valores iniciais, é sugerida a utilização de uma grade de possíveis valores.

5.4 Divisão dos Nós

Há aqui um particular interesse na hipótese relacionada a significância da divisão do nó inicial (raiz).

Ao reparametrizar o modelo definido por (5-1)–(5-2) como:

$$y_t = \phi_0 + \lambda_0 G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \varepsilon_t, \tag{5-7}$$

onde $\phi_0 = \beta_2$ and $\lambda_0 = \beta_1 - \beta_2$, é obtida uma representação mais parcimoniosa do modelo STR-Tree para a árvore mais simples[2].

Com o objetivo de testar a significância da primeira divisão, a hipótese nula conveniente é $\mathcal{H}_0 : \gamma_0 = 0$, contra a alternativa $\mathcal{H}_a : \gamma_0 > 0$. Uma hipótese nula equivalente é $\mathcal{H}'_0 : \lambda_0 = 0$.

Entretanto, fica claro em (5-7) que sob $\mathcal{H}_0$, os parâmetros de perturbação (*nuisance*) λ_0 and c_0 podem assumir diferentes valores sem que a função de verossimilhança seja modificada. Isto cria um problema de identificação cuja primeira solução foi discutida em [30]. Para este tópico recomenda-se também o trabalho de [31].

Para este problema, adota-se a solução proposta em [64] [3], que aproxima a função $G(\cdot)$ por uma expansão de Taylor de 3a. ordem em torno de $\gamma = 0$. Após algumas manipulações algébricas, chega-se a:

$$y_t = \alpha_0 + \alpha_1 x_{s_0,t} + \alpha_2 x^2_{s_0,t} + \alpha_3 x^3_{s_0,t} + e_t, \tag{5-8}$$

onde α_i, $i = 0, 1, 2, 3$, é um parâmetro que está em função de γ_0, c_0, ϕ_0, e λ_0, $e_t = \varepsilon_t + \lambda_0 R(\mathbf{x}_t; s_0, \gamma_0, c_0)$, e $R(\mathbf{x}_t; s_0, \gamma_0, c_0)$ é o termo remanescente na expansão.

Sendo assim, a hipótese nula de interesse passa a ser :

[2] Torna-se fácil notar que (5-7) é um caso particular de uma rede neural com uma única camada escondida ([53]).

[3] veja também [85]

$$\mathcal{H}_0 : \alpha_i = 0, \ i = 1, 2, 3. \tag{5-9}$$

Note que sob $\mathcal{H}_0$ o resto da expansão de Taylor desaparece e $e_t = \varepsilon_t$, de modo que as propriedades do erro permanecem inalteradas sob a hipótese nula, gerando condições propícias para o uso de inferência assintótica .

Finalmente, pode-se apontar que (5-8) pode ser visto como resultado de uma aproximação local para a função de log-verossimilhança, que, para a t-ésima observação assume a forma:

$$l_t = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left\{y_t - \alpha_0 - \alpha_1 x_{s_0,t} - \alpha_2 x_{s_0,t}^2 - \alpha_3 x_{s_0,t}^3\right\}^2. \tag{5-10}$$

Neste ponto, levanta-se uma hipótese adicional, para acompanhar as Hipóteses (2)–(4).

Hipótese 5 $E|x_{s_0 t}|^\delta < \infty, \forall\ s_0 \in \mathbb{S}$, *para algum* $\delta > 6$.

Isto possibilita enunciar alguns resultados que são bem conhecidos na literatura econométrica.

Teorema 5.4 *Sob* $\mathcal{H}_0 : \gamma_0 = 0$ *e Hipóteses (2)–(5), a estatística do teste tipo ML*

$$ML = \frac{1}{\widehat{\sigma}^2}\sum_{t=1}^{T}\widehat{\varepsilon}_t\boldsymbol{\nu}_t'\left\{\sum_{t=1}^{T}\boldsymbol{\nu}_t\boldsymbol{\nu}_t' - \sum_{t=1}^{T}\boldsymbol{\nu}_t\mathbf{h}_t'\left(\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t'\right)^{-1}\sum_{t=1}^{T}\mathbf{h}_t\boldsymbol{\nu}_t'\right\}^{-1}\sum_{t=1}^{T}\boldsymbol{\nu}_t\widehat{\varepsilon}_t, \tag{5-11}$$

onde $\widehat{\varepsilon}_t = y_t - \widehat{\beta}_0$, *é o conjunto de resíduos estimados sob a hipótese nula,* $\widehat{\sigma}^2 = (1/T)\sum_{t=1}^{T}\widehat{\varepsilon}_t^{\,2}$, $\mathbf{h}_t = 1$,*e* $\boldsymbol{\nu}_t = \left(x_{s_0 t}, x_{s_0 t}^2, x_{s_0 t}^3\right)'$, *tem distribuição assintótica* χ^2 *com* 3 *graus de liberdade.*

Comentário 2 *Note que, sob* $\mathcal{H}_0$, $\widehat{\beta}_0 = \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} E(y_t)$.

Até este ponto, o modelo STR-Tree tem sido interpretado como um caso particular do modelo STR em [41], e a estratégia de testar a divisão do nó raiz corresponde ao teste de linearidade onde o modelo linear em questão é o constante global.

Entretanto, a idéia chave neste trabalho é considerar este procedimento de teste dentro de uma abordagem mais complexa. Para um modelo mais complexo, considere que a hipótese nula (5-9) tenha sido rejeitada e os parâmetros do modelo STR-Tree mais simples tenham sido consistentemente estimados.

Um caminho natural, dentro da estruturação por árvores, é considerar a hipótese de erro de especificação através da formulação de um novo modelo

que divide um dentre os dois nós criados; seja este, por exemplo, o nó filhote esquerdo, o que irá constituir o seguinte modelo (ver Figura 3.5):

$$\begin{aligned} y_t =& H(\mathbf{x}_t;\boldsymbol{\psi}) + \varepsilon_t \\ =& \{\beta_3 G(\mathbf{x}_t; s_1, \gamma_1, c_1) + \beta_4 [1 - G(\mathbf{x}_t; s_1, \gamma_1, c_1)]\} G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \\ & \beta_2 [1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)] + \varepsilon_t. \end{aligned} \tag{5-12}$$

Portanto, re-escrevendo (5-12) como

$$\begin{aligned} y_t =& [\phi_1 + \lambda_1 G(\mathbf{x}_t; s_1, \gamma_1, c_1)] G(\mathbf{x}_t; s_0, \gamma_0, c_0) + \\ & \beta_2 [1 - G(\mathbf{x}_t; s_0, \gamma_0, c_0)] + \varepsilon_t, \end{aligned} \tag{5-13}$$

onde $\phi_1 = \beta_3$ and $\lambda_1 = \beta_3 - \beta_4$, a hipótese nula conveniente nesta ocasião é $\mathcal{H}_0 : \gamma_1 = 0$.

Entretanto, sob esta hipótese nula, o modelo (5-13) não pode ser consistentemente estimado em função dos parâmetros de perturbação λ_1 e c_1. Para resolver este problema de identificação, o procedimento é o mesmo adotado anteriormente, ou seja, aproxima-se a função $G(\cdot)$ por sua expansão em série de Taylor de 3a. ordem em torno de $\mathcal{H}_0$.

Após alguns cálculos, é obtido o seguinte resultado a partir da expansão:

$$\begin{aligned} y_t =& \alpha_0 + \alpha_1 G(x_{s_0 t}; \gamma_0, c_0) + \alpha_2 G(x_{s_0 t}; \gamma_0, c_0) x_{s_1 t} + \\ & \alpha_3 G(x_{s_0 t}; \gamma_0, c_0) x_{s_1 t}^2 + \alpha_4 G(x_{s_0 t}; \gamma_0, c_0) x_{s_1 t}^3 + e_t, \end{aligned} \tag{5-14}$$

onde $e_t = \varepsilon_t + R(\mathbf{x}_t; s_1, \gamma_1, c_1)$; $R(\mathbf{x}_t; s_1, \gamma_1, c_1)$ sendo o resto.

A decisão pela divisão do nó corresponde à rejeição da seguinte hipótese nula

$$\mathcal{H}_0 : \alpha_i = 0, \ i = 2, 3, 4. \tag{5-15}$$

A estatística de teste é dada por(5-11) com

$$\mathbf{h}_t = \left.\frac{\partial H(\mathbf{x}_t;\boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}\right|_{\mathcal{H}_0} = \left(1, \ G(x_{s_0 t}; \widehat{\gamma}_0, \widehat{c}_0), \ \left.\frac{\partial G(x_{s_0 t}; \gamma_0, c_0)}{\partial \gamma_0}\right|_{\mathcal{H}_0}, \ \left.\frac{\partial G(x_{s_0 t}; \gamma_0, c_0)}{\partial c_0}\right|_{\mathcal{H}_0}\right)' \tag{5-16}$$

e

$$\boldsymbol{\nu}_t = \left(G\left(x_{s_0t};\widehat{\gamma}_0,\widehat{c}_0\right)x_{s_1t},\, G\left(x_{s_0t};\widehat{\gamma}_0,\widehat{c}_0\right)x_{s_1t}^2,\, G\left(x_{s_0t};\widehat{\gamma}_0,\widehat{c}_0\right)x_{s_1t}^3\right)'. \quad (5\text{-}17)$$

A partir da suposição da normalidade dos erros, a matriz de informação é bloco diagonal e assim podemos assumir que a variância do erro é fixa. Se utilizada a versão F do teste do tipo ML, a partir do vetor escore, sob a hipótese nula, pode ser visto que a estatística de decisão pode ser calculada de acordo com os seguintes passos :

1. Estimar o modelo STR-Tree sob a hipótese nula de que $\mathcal{H}_0$ e calcular os resíduos $\widehat{\varepsilon}_t$. Computar a soma dos resíduos quadráticos $SSR_0 = \sum_{t=1}^{T}\widehat{\varepsilon}_t^2$.

2. Fazer a regressão de $\widehat{\varepsilon}_t$ em $\mathbf{h}_t$ e $\boldsymbol{\nu}_t$. Calcular a soma dos resíduos quadráticos obtidos a partir desta regressão, $SSR_1 = \sum_{t=1}^{t}\widehat{u}_t^2$.

3. Calcular a estatística χ^2

$$LM_\chi = T\frac{SSR_0 - SSR_1}{SSR_0}, \quad (5\text{-}18)$$

ou a versão F do teste

$$LM_F = \frac{\left(SSR_0 - SSR_1\right)/3}{SSR_1/(T-7)}, \quad (5\text{-}19)$$

onde T é o tamanho de amostra.

Sob a hipótese nula, LM_χ é assintoticamente distribuído de acordo com a distribuição χ^2 com 3 graus de liberdade e LM_F tem distribuição assintótica F com 3 e $T-7$ graus de liberdade.

Daqui para a frente, a idéia é executar uma seqüência de testes do tipo ML para crescer a árvore no mesmo formato apresentado acima. A forma geral da estatística de teste, quando testa-se um modelo com j nós contra a alternativa de que o modelo tem $j+1$ nós, é dada por:

$$LM = \frac{\left(SSR_0 - SSR_1\right)/3}{SSR_1/[T-(p+3)]}, \quad (5\text{-}20)$$

onde p é o número total de elementos no vetor $\mathbf{h}_t$.

A estratégia de modelagem é descrita nas próximas seções.

O Ciclo de Modelagem a Partir da Raiz (profundidade 0)

A decisão para divisão da raiz é baseada nos seguintes passos.

1. Para cada variável explanatória, aplicar o teste do tipo ML descrito acima e selecionar a variável x_{s_0t} que gera o menor p-valor abaixo de um nível de significância pré-especificado α.

2. Condicionalmente à escolha de s_0, estima-se o vetor de parâmetros $\boldsymbol{\psi} = (\gamma_0, c_0, \beta_1, \beta_2)'$ pela minimização dos quadrados não-linear concentrada.

3. Avalia-se o modelo estimado pelos seguintes testes de hipóteses (condicional em γ_0 e c_0)

$$\begin{aligned} &\mathcal{H}_{01}: \beta_1 = 0 \\ &\mathcal{H}_{02}: \beta_2 = 0 \\ &\mathcal{H}_{03}: \beta_1 - \beta_2 = 0 | \beta_1, \beta_2 \neq 0 \end{aligned} \tag{5-21}$$

contra alternativas bilaterais.

Se pelo menos uma das hipóteses avaliadas em (5-21) não rejeitar a hipótese nula, o ciclo retorna para o estágio de especificação e a próxima variável de divisão ligada ao *ranking* de p-valores é selecionada. No caso de todas as variáveis não produzirem uma divisão estatisticamente significante, a raiz é declarada como nó terminal e o modelo constante global é selecionado como melhor modelo. Caso contrário, dois nós filhotes são gerados, compondo a primeira profundidade da árvore.

Ciclo de Modelagem a partir da 1a. Profundidade

Após a árvore ter começado a crescer a partir da raiz, a primeira profundidade é criada e o ciclo tem continuidade colocando como hipótese alternativa ao atual modelo a divisão de um dentre os dois nós terminais existentes. A hipótese nula neste teste está relacionada ao modelo linear condicional e a alternativa contém a inclusão de um termo não-linear que é responsável pela partição do nó.

De agora em diante, além da seleção de uma variável de divisão, deve também ser selecionado qual nó deve ser dividido em primeiro lugar [4].

[4] Como os testes são feitos de forma condicional aos parâmetros previamente estimados, decidir qual o primeiro nó a ser dividido torna-se importante.

1. Para cada combinação do índice da variável de divisão em $\mathbb{S} = \{1, 2, \ldots, m\}$ e o número do nó em $\mathbb{D}_1 = \{1, 2\}$, aplica-se o teste do tipo ML e seleciona-se os índices $j_1 \in \mathbb{D}_1$ e $s_{j_1} \in \mathbb{S}$ que geram o menor p-valor abaixo de um nível de significância pré-estabelecido.

2. Estima-se os parâmetros do modelo.

3. O modelo é avaliado através dos testes das seguintes hipóteses:

$$\begin{aligned} &\mathcal{H}_{01} : \beta_{2j_1+1} = 0 \\ &\mathcal{H}_{02} : \beta_{2j_1+2} = 0 \\ &\mathcal{H}_{03} : \beta_{2j_1+1} - \beta_{2j_1+2} = 0 | \beta_{2j_1+1}, \beta_{2j_1+2} \neq 0 \end{aligned} \tag{5-22}$$

Caso não seja encontrada significância nos testes listados acima, o modelo deve ser re-especificado pela escolha de uma nova combinação de nó e variável de divisão. Se a divisão é aceita, então o ciclo retorna para o primeiro passo pela aplicação do teste do tipo ML para testar o modelo com 3 nós teminais contra a alternativa que divide o nó $j_2 \in \mathbb{D}_1 - \{j_1\}$. A segunda profundidade estará completa quando ambos os nós j_1 e j_2 produzirem divisões significativas. No caso em que j_1 é o único dos nós a gerar filhotes, a segunda profundidade estará composta por dois nós cujos números serão: $2j_1 + 1$ and $2j_1 + 2$. Caso não haja divisões significativas, o processo de crescimento da árvore é encerrado.

Ciclo de Modelagem a partir da $k-$ésima profundidade

A execução do algoritmo em uma profundidade genérica $k > 0$ é análoga aos desenvolvimentos já mostrados.

1. Aplicar o teste ML para todas as combinações de variáveis de divisão e nós dentro do conjunto $\mathbb{D}_k$ que contém todos os números dos nós filhotes que compõe a $k-$ésima profundidade. Note que $\mathbb{D}_k \subseteq \left\{2^k - 1, 2^k, \ldots, 2^{k+1} - 2\right\}$.

2. Selecione $j_1 \in \mathbb{D}_k$ e $s_{j_k} \in \mathbb{S}$ pelo *ranking* dos p-valores abaixo do nível de significância, obtidos pelo teste do tipo ML.

3. Estimação dos parâmetros do modelo.

4. Avaliação do modelo pela verificação das estatísticas t das constantes dentro dos nós criados e a significância da diferença entre elas.

O ciclo nesta profundidade é executado iterativamente pelo teste, e se necessário, dividindo os nós de acordo com a sequência:

$$
\begin{aligned}
&j_2 \in \mathbb{D}_1 - \{j_1\} \\
&j_3 \in \mathbb{D}_1 - \{j_1, j_2\} \\
&j_4 \in \mathbb{D}_1 - \{j_1, j_2, j_3\} \\
&\dots
\end{aligned}
$$

Ao atingir um ponto em que não haja mais divisões significativas, o algoritmo dirige-se para a profundidade $(k + 1)$. O encerramento do ciclo ocorrerá quando uma determinada profundidade não gerar mais nós filhotes.

5.4.1 Testes Sequenciais

Para alcançar a especificação final do modelo estruturado por árvores, é executada uma sequência de n testes de hipóteses, correlacionados, do tipo ML, na qual n é uma variável aleatória. Durante esta seqüência, a decisão prejudicial a ser tomada, de acordo com o princípio de que a complexidade da árvore é função exclusiva do seu número de nós terminais, é decidir equivocadamente pela divisão do nó.

Devido à multiplicidade originada a partir dos repetidos testes de significância, deve ser criada uma estratégia para controlar o erro (geral) do tipo I sob o risco de superestimar a quantidade de resultados significativos, ou seja, reportar um número de divisões significativas acima daquelas que deveriam ter realmente ocorrido.

Para remediar esta situação, adota-se aqui o seguinte procedimento. Para o $n-$ésimo teste na sequência, caso este seja aplicado na $d-$ésima profundidade, o nível de significância é fixado em $\alpha(d, n) = \frac{\alpha}{n^d}$.

Na raiz, $(d = 0)$, desta forma, o primeiro teste é aplicado para verificar a significância da divisão sob um nível α. Se a hipótese nula é rejeitada, então o segundo teste $(n = 2)$ é aplicado na primeira profundidade, o que implica em fixar o nível de significância em $\alpha/2$.

Então, se a árvore cresce completando os possíveis nós em todas as profundidades, o nível de significância evolui do seguinte modo: $\alpha/3$, $\alpha/4^2$, $\alpha/5^2$, $\alpha/6^3$, $\alpha/7^3$, $\alpha/8^4$, $\alpha/9^4$, etc...

A Figura 5.2 exibe um exemplo hipotético da evolução do nível de significância durante o processo de construção da árvore. As setas na Figura 5.2

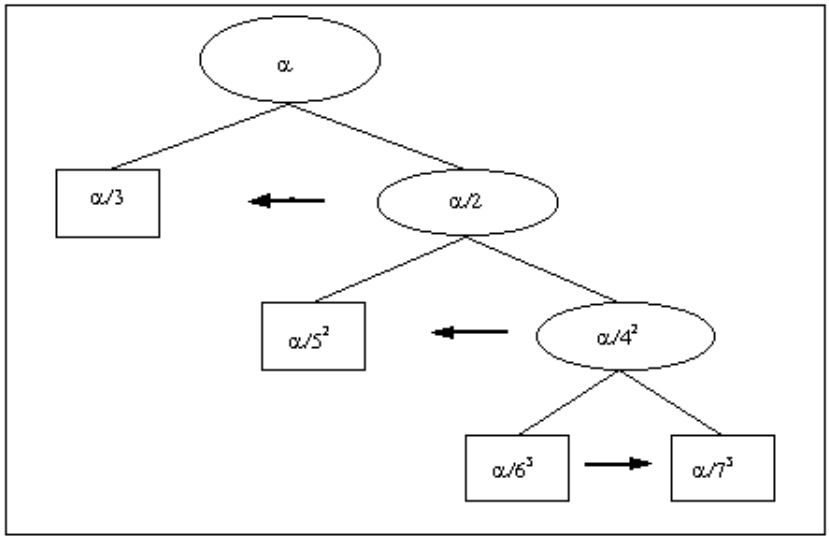

Figura 5.2: Nível de significância durante a seqüência de testes para divisão dos nós - um exemplo hipotético

mostram a ordem em que os nós são testados para divisão em cada profundidade da árvore.

Ao forçar o teste a ser mais rigoroso em grandes profundidades, cria-se uma estratégia que diminui a importância de utilizar técnicas de podagem a posteriori (*post-prunning*).

Existem outras propostas para controlar o tamanho geral do teste, para isto sugere-se olhar [52, 14, 12, 13, 7, 8]. Entretanto, pelos experimentos realizados nesta tese, a metodologia descrita acima parece funcionar satisfatoriamente e a comparação com outras diferentes técnicas de redução do nível nominal do teste não serão aplicadas neste trabalho.

Na prática, diferentes metodologias podem ser testadas e possíveis arquiteturas diferentes podem ser comparadas pelo desempenho fora-da-amostra.

5.4.2 Dados Categóricos

A princípio, o desenvolvimento das seções anteriores não considera a possibilidade de algumas variáveis de divisão serem categóricas. Entretanto, a extensão para variáveis categóricas é direta. A idéia principal é substituir os modelos constantes dentro das folhas por uma regressão linear em um conjunto de variáveis indicadoras representando os dados categóricos.

Faça $\mathbf{x}_t = (\mathbf{z}_t', \mathbf{w}_t')'$, onde $\mathbf{z}_t$ é um vetor de variáveis categóricas e $\mathbf{w}_t$ é um vetor de variáveis contínuas. Adicionalmente, faça $\mathbf{D}_t(\mathbf{z}_t)$ ser um vetor de

variáveis indicadoras representando o vetor categórico $\mathbf{z}_t$. Neste caso o modelo l (5-3) pode ser re-escrito como:

$$y_t = H\left(\mathbf{x}_t; \boldsymbol{\psi}\right) + \varepsilon_t = \sum_{k=1}^{K} \boldsymbol{\beta}'_{K+i-1}\mathbf{D}_t(\mathbf{z}_t)B_k(\mathbf{w}_t; \boldsymbol{\theta}_k) + \varepsilon_t. \qquad (5\text{-}23)$$

Esta abordagem é similar à utilizada na literatura sobre modelos STR para lidar com a presença de regressores do tipo *dummy*.

5.5 Experimento de Monte Carlo

Um experimento de Monte Carlo foi planejado com dois objetivos. O primeiro é estudar as propriedades dos estimadores MQNL em pequenas amostras, e o segundo é investigar o desempenho de 3 diferentes algoritmos para construção de uma árvore:

CART: Aqui foi utilizada a forma mais tradicional do algoritmo CART. Esta consiste em crescer a árvore de maior tamanho possível, utilizando como regra de parada o mínimo de 5 observações por nó terminal, e depois realizar um podagem com a regra 1-SE com erros estimados por validação cruzada com 10-dobras.

STR-Tree/LM: Como descrito nas seções anteriores, esta estratégia utiliza uma sequência de testes do tipo ML para selecionar simultaneamente o nó a ser dividido e a variável de divisão. Esta estratégia de especificação não necessita de procedimento de podagem e o controle do erro global é feito pela redução do tamanho do teste durante o crescimento da árvore.

STR-Tree/CV: Na tentativa de utilizar uma estratégia semelhante ao CART, realiza-se um experimento de validação cruzada com 10-dobras para selecionar a variável de divisão que minimiza o Erro Médio Quadrático global avaliado fora-da-amostra. Quando o EQM mais um erro padrão é superior ao encontrado na divisão anterior, o nó é declarado como terminal.

Foram simuladas duas arquiteturas com um número pequeno de folhas, ilustradas na Figura 5.3. Pela seleção destas duas arquiteturas que apresentam complexidade maior do que aquela que contém apenas dois nós, foram simulados modelos para diferentes combinações dos parâmetros de suavidade dentro dos nós terminais. Desta forma, 4 modelos foram simulados para a Arquitetura I que

contém 3 nós terminais e dois modelos foram simulados para a Arquitetura II que contém 4 nós terminais.

Basicamente, considera-se na Tabela 5.1 dois tipos de divisão, suave ($\gamma_i = 0.5$) e rígida ($\gamma_i = 5$), que foram combinadas em diferentes sequências durante o processo de construção da árvore simulada. O Modelo l 1.1, por exemplo, é obtido a partir de duas divisões suaves consecutivas, e o Modelo 1.4 aplica uma divisão suave nos dados da raiz, seguida de uma divisão rígida.

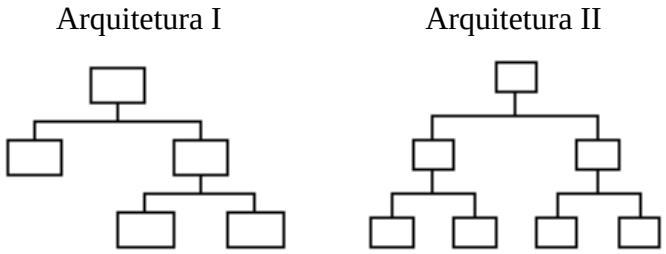

Figura 5.3: Arquiteturas Simuladas

Tabela 5.1: Suavidade das Divisões nas Simulações do STR-Tree

	Modelo	Primeira Divisão	Segunda Divisão	Terceira Divisão
Arquitetura I	1.1	$\gamma_0 = 0.5$	$\gamma_2 = 0.5$	—
(3 folhas)	1.2	$\gamma_0 = 5$	$\gamma_2 = 5$	—
	1.3	$\gamma_0 = 5$	$\gamma_2 = 0.5$	—
	1.4	$\gamma_0 = 0.5$	$\gamma_2 = 5$	—
Arquitetura II	2.1	$\gamma_0 = 0.5$	$\gamma_1 = 0.5$	$\gamma_2 = 0.5$
(4 folhas)	2.2	$\gamma_0 = 5$	$\gamma_1 = 5$	$\gamma_2 = 5$

Foram feitas 1000 replicações para cada modelo com tamanhos de amostra $T = 150$ e $T = 500$, numa tentativa de representar pequenas e grandes amostras. Como a preocupação principal foi direcionada aos efeitos das escolhas para o parâmetro de suavidade, não houve muita variação na escolha dos parâmetros associados aos modelos constantes dentro dos nós. Três preditores não correlacionados foram utilizados como candidatos à variável de divisão $x_1 \sim \mathrm{N}(10, 2.56)$; $x_2 \sim \mathrm{N}(90, 9)$; e $x_3 \sim \mathrm{N}(25, 4)$.

O erro foi definido comos $\varepsilon_t \sim \mathrm{N}(0, 1)$. Desde que o parâmetro de suavidade é dependente da escala, o argumento da função logística foi padronizado, através da divisão pelo desvio padrão da variável de divisão . Os outros parâmetros foram fixados de acordo com os valores na Tabela 5.2.

Como mostrado na Tabela 5.2,os parâmetros de locação foram escolhidos estratégicamente nos seus valores medianos para simulações da Arquitetura II. O objetivo foi proporcionar uma boa quantidade de informação dentro dos nós

Tabela 5.2: Parâmetros do Modelo STR-Tree Simulado

	Arquitetura I	Arquitetura II
Constantes dentro dos nós	$\beta_1 = 6$ $\beta_5 = 1.8$; $\beta_6 = -1.5$	$\beta_3 = 6$; $\beta_4 = 3.2$ $\beta_5 = 1.8$; $\beta_6 = -1.5$
Parâmetros de Locação	$c_0 = 83$; $c_2 = 10$	$c_0 = 90$; $c_1 = 10$; $c_2 = 25$
Indices das Variáveis de Divisão	$s_0 = 2$; $s_2 = 1$	$s_0 = 2$; $s_1 = 1$; $s_2 = 3$

gerados. A única preocupação na escolha das constantes dentro dos nós foi produzir diferentes modelos locais.

As combinações apresentadas acima proporcionaram diferentes relações entre a variável resposta e o conjunto de variáveis explanatórias

Ao contrário do CART, que ajusta um histograma multidimensional aos dados, o modelo STR-Tree ajusta uma superfície. A diferença entre os modelos para Arquitetura I pode ser vista na Figura 5.4, que traz a superfície de resposta para cada uma das árvores simuladas.

Quando todas as divisões são rígidas, tal como no Modelo 1.2, a superfície se parece com um histograma bivariado. Por outro lado, uma sequência de divisões extremamente suaves (Modelo 1.1) produz uma relação entre a resposta e os regressores que é praticamente linear.

Estimação dos Parâmetros

Nesta seção, são apresentados resultados empíricos obtido com o uso de MQNL nos modelos simulados. Os resultados são descritos através de estatísticas descritivas tais como a média e a mediana para avaliar a tendência central. Para avaliar a variabilidade das estimativas foram escolhidas duas medidas; o desvio padrão, e como uma alternativa mais robusta, desvio absoluto mediano em torno da mediana (DAM).

$$DAM(\widehat{\psi}) = \text{mediana}\left(\left|\widehat{\psi} - \text{mediana}(\widehat{\psi})\right|\right). \tag{5-24}$$

A estimação do parâmetro de suavidade γ resulta em *outliers* e valores extremos para algumas simulações, consequentemente a média amostral das estimativas é fortemente afetada por estes.

As Tabelas 5.3 e 5.4 mostram que o parâmetro γ, para algumas das replicações, é fortemente superestimado quando $T = 150$. Nestes casos, a mediana aparenta ser uma medida mais robusta de tendência central.

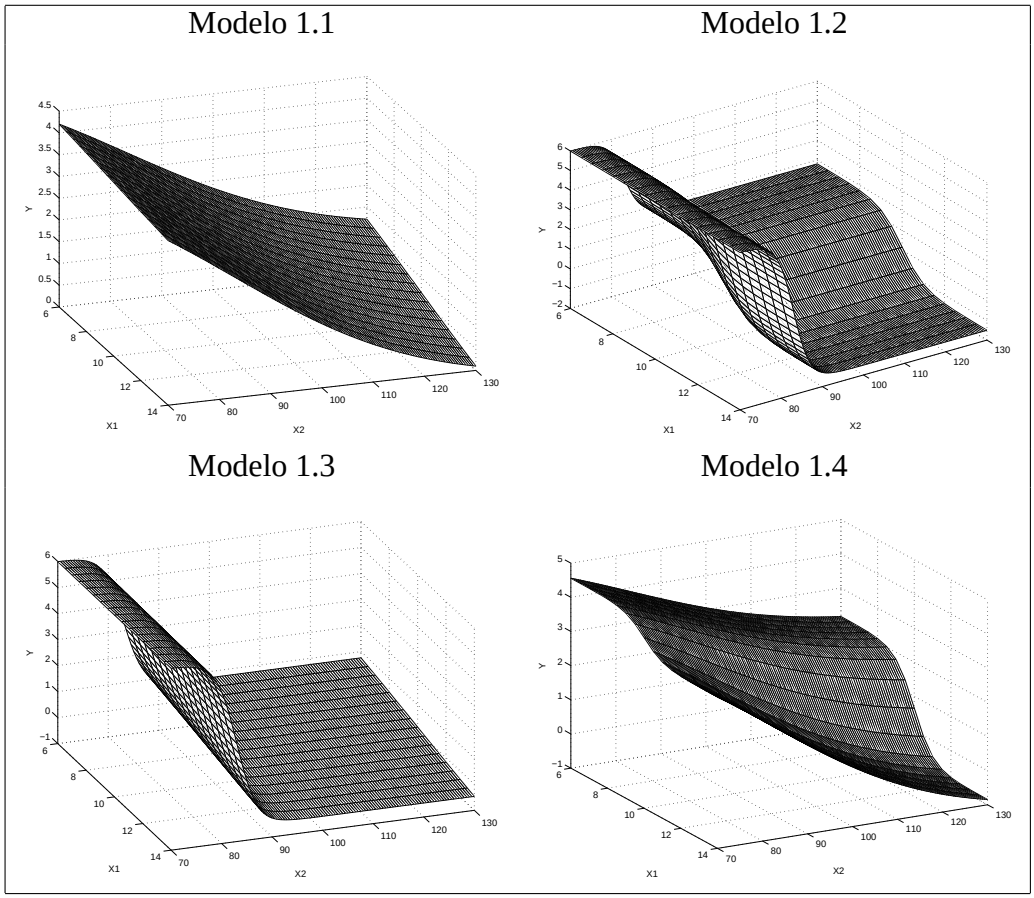

Figura 5.4: Características Geométricas dos Modelos Simulados (Arquitetura I)

Por outro lado, tal problema não ocorre com o parâmetro de locação cuja média amostral e mediana estão próximos do valor verdadeiro. Não obstante este fato, a variabilidade do parâmetro de locação aumenta sempre que ocorre uma divisão suave. Como conseqüência disto, as estimativas dos parâmetros dentro dos nós são também afetadas, principalmente em amostras pequenas.

Assim, tal como aconteceu com o Modelo 1.3, a média amostral e a mediana para os modelos locais desviam dos valores populacionais.

As linhas que compõe a grade na Figura 5.5 foram traçadas perpendicularmente aos verdadeiros valores dos parâmetros ($c_2 = 10$ and $\beta_4 = -1.5$). Fica evidente que desvios positivos em relação ao parâmetro de locação, gerando uma quantidade menor de observações alocadas dentro do nó direito, fazem com que ocorram valores extremos nas estimativas das constantes dentro dos nós.

Exceto pelos parâmetro de suavidade, as estimativas foram, em geral, mais precisas para árvores simuladas com divisões rígidas. Ao combinar divisões rígidas com divisões suaves, os resultados apontaram que uma divisão suave precedendo uma divisão rígida produz resultados com maior acurácia, do que a situação contrária. Nesta situação uma justificativa plausível é a de que sobram

Tabela 5.3: Estatísticas Descritivas para a Estimação na Arquitetura I

Modelo 1.1		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (0.5)	0.518	0.112	0.502	0.066	0.503	0.055	0.498	0.036
$\hat{c}_0$ (83)	82.988	0.476	83.002	0.313	83.010	0.236	83.015	0.150
$\hat{\gamma}_2$ (0.5)	25183	207.942	0.570	0.268	0.533	0.178	0.522	0.113
$\hat{c}_2$ (10)	10.020	2.255	10.036	0.694	10.032	0.812	10.006	0.372
$\hat{\beta}_1$ (6)	6.016	0.364	6.007	0.229	6.004	0.173	5.996	0.113
$\hat{\beta}_5$ (1.8)	2.187	1.531	1.734	0.526	1.895	0.567	1.766	0.252
$\hat{\beta}_6$ (-1.5)	-1.915	1.510	-1.452	0.512	-1.623	0.630	-1.472	0.250
Modelo 1.2		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (5)	17.059	60.519	5.254	2.297	6.190	9.580	5.154	1.126
$\hat{c}_0$ (83)	83.035	0.183	83.019	0.097	83.008	0.071	83.002	0.042
$\hat{\gamma}_2$ (5)	35.672	319.697	5.581	1.642	11.260	153.725	5.158	0.767
$\hat{c}_2$ (10)	10.002	0.099	10.004	0.066	9.998	0.051	9.997	0.035
$\hat{\beta}_1$ (6)	6.012	0.189	6.013	0.128	5.996	0.106	5.998	0.072
$\hat{\beta}_5$ (1.8)	1.789	0.159	1.792	0.105	1.799	0.088	1.801	0.056
$\hat{\beta}_6$ (-1.5)	-1.501	0.161	-1.497	0.102	-1.501	0.087	-1.496	0.058
Modelo 1.3		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (5)	10.917	21.949	5.288	1.693	5.852	9.369	5.100	0.870
$\hat{c}_0$ (83)	83.006	0.146	82.998	0.073	82.999	0.061	82.998	0.040
$\hat{\gamma}_2$ (0.5)	16.131	126.012	0.542	0.238	0.526	0.171	0.520	0.107
$\hat{c}_2$ (10)	10.062	2.1281	9.969	0.707	10.003	0.964	10.007	0.368
$\hat{\beta}_1$ (6)	6.009	0.193	6.007	0.126	5.999	0.102	5.998	0.064
$\hat{\beta}_5$ (1.8)	2.204	1.420	1.766	0.509	1.953	0.739	1.785	0.243
$\hat{\beta}_6$ (-1.5)	-1.955	1.595	-1.441	0.464	-1.653	0.732	-1.483	0.246
Modelo 1.4		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (0.5)	0.527	0.145	0.505	0.0709	0.506	0.066	0.503	0.043
$\hat{c}_0$ (83)	83.045	0.513	83.023	0.342	83.011	0.277	83.020	0.183
$\hat{\gamma}_2$ (5)	45.670	386.809	5.402	1.779	9.213	110.411	5.077	0.741
$\hat{c}_2$ (10)	10.002	0.111	10.005	0.072	9.999	0.051	9.999	0.032
$\hat{\beta}_1$ (6)	6.004	0.357	5.984	0.223	6.000	0.188	5.994	0.123
$\hat{\beta}_5$ (1.8)	1.778	0.182	1.789	0.117	1.791	0.096	1.795	0.066
$\hat{\beta}_6$ (-1.5)	-1.511	0.219	-1.505	0.145	-1.503	0.116	-1.500	0.078

Tabela 5.4: Estatísticas Descritivas para a Estimação na Arquitetura II

Modelo 2.1		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (0.5)	0.693	3.261	0.510	0.075	0.508	0.068	0.504	0.043
$\hat{c}_0$ (90)	90.017	0.542	89.998	0.380	89.999	0.305	89.994	0.198
$\hat{\gamma}_1$ (0.5)	46.594	397.130	0.805	0.531	17.417	253.162	0.557	0.188
$\hat{c}_1$ (10)	10.104	2.135	10.095	1.028	9.962	1.311	9.942	0.578
$\hat{\gamma}_2$ (0.5)	11.897	171.624	0.549	0.197	0.535	0.173	0.512	0.100
$\hat{c}_2$ (25)	24.997	1.953	25.031	0.781	24.990	0.710	24.997	0.358
$\hat{\beta}_3$ (6)	6.104	1.215	5.730	0.479	6.132	0.762	5.956	0.344
$\hat{\beta}_4$ (3.2)	3.045	1.261	3.429	0.445	3.102	0.729	3.271	0.323
$\hat{\beta}_5$ (1.8)	2.061	1.155	1.773	0.372	1.854	0.437	1.797	0.201
$\hat{\beta}_6$ (-1.5)	-1.777	1.091	-1.520	0.423	-1.555	0.451	-1.491	0.205
Modelo 2.2		$T = 150$				$T = 500$		
	Média	Desv.Pad.	Mediana	DAM	Média	Desv.Pad.	Mediana	DAM
$\hat{\gamma}_0$ (5)	70.192	1276.098	5.530	2.998	25.373	238.576	5.080	1.389
$\hat{c}_0$ (90)	90.009	0.238	90.002	0.130	90.003	0.116	89.997	0.065
$\hat{\gamma}_1$ (5)	104.765	527.811	6.993	3.932	367.216	4863.138	5.471	1.470
$\hat{c}_1$ (10)	9.997	0.157	9.999	0.094	10.005	0.082	10.006	0.056
$\hat{\gamma}_2$ (5)	76.126	553.641	6.700	3.596	55.747	323.296	5.207	1.261
$\hat{c}_2$ (25)	24.995	0.182	24.999	0.103	25.001	0.085	24.999	0.053
$\hat{\beta}_3$ (6)	6.004	0.218	6.004	0.132	5.990	0.124	5.988	0.084
$\hat{\beta}_4$ (3.2)	3.210	0.209	3.216	0.139	3.213	0.115	3.216	0.073
$\hat{\beta}_5$ (1.8)	1.790	0.194	1.782	0.125	1.789	0.099	1.794	0.067
$\hat{\beta}_6$ (-1.5)	-1.494	0.204	-1.487	0.128	-1.492	0.116	-1.493	0.079

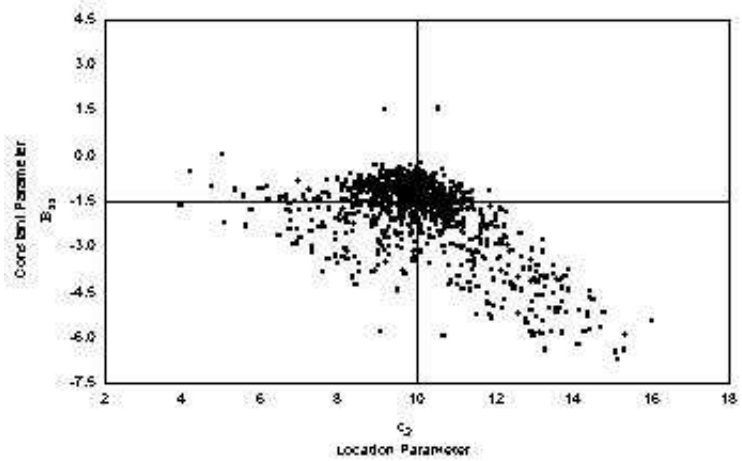

Figura 5.5: Diagrama de dispersão entre as estimativas do parâmetro de locação c_2 e da constante dentro do nó para o modelo 1.1 (T=150)

mais pontos para serem modelados após a primeira divisão.

Finalmente, um importante resultado obtido a partir deste experimento de Monte Carlo foi a convergência das estimativas de MQNL para os valores verdadeiros quando aumentou-se o tamanho da amostra. Estes resultados eram esperados conforme as propriedades descritas na seção anterior.

Especificação da Arquitetura da Árvore por Diferentes Algoritmos

Mostra-se na Tabela 5.5 e Tabela 5.6, o desempenho dos 3 algoritmos citados na identificação de modelos STR-Tree simulados. Estes resultados são mostrados de forma mais detalhada em C.1.

Tabela 5.5: Percentual de Especificações Corretas para Árvore Simuladas na Arquitetura I

		$T = 150$	
Parâmetro de Suavidade	CART	STR-Tree/LM	STR-Tree/CV
γ_0=0.5 γ_2=0.5	7.7%	34.7%	6.4%
γ_0=5 γ_2=5	8.4%	98.4%	89.9%
γ_0=5 γ_2=0.5	16.4%	85.4%	42.5%
γ_0=0.5 γ_2=5	37.8%	45.8%	38.4%
		$T = 500$	
Parâmetro de Suavidade	CART	STR-Tree/LM	STR-Tree/CV
γ_0=0.5 γ_2=0.5	23%	84.2%	15.1%
γ_0=5 γ_2=5	0%	97.8%	96.3%
γ_0=5 γ_2=0.5	0.1%	99.1%	80.6%
γ_0=0.5 γ_2=5	3.5%	6.1%	11.4%

Tabela 5.6: Percentual de Especificações Corretas para Árvore Simuladas na Arquitetura II

		$T = 150$	
Parâmetro de Suavidade	CART	STR-Tree/LM	STR-Tree/CV
$\gamma_0 = 0.5\ \gamma_1 = 0.5\ \gamma_2 = 0.5$	0.8%	4%	0.6%
$\gamma_0 = 5\ \gamma_1 = 5\ \gamma_2 = 5$	25.9%	98.3%	76.7%
		$T = 500$	
Parâmetro de Suavidade	CART	STR-Tree/LM	STR-Tree/CV
$\gamma_0 = 0.5\ \gamma_1 = 0.5\ \gamma_2 = 0.5$	4.3%	61.1%	1.3%
$\gamma_0 = 5\ \gamma_1 = 5\ \gamma_2 = 5$	0%	98%	94.8%

Quando todas as partições envolveram apenas divisões rígidas, o modelo STR-Tree produziu um percentual de especificações corretas acima de 95%, independente da arquitetura simulada e, quando $T = 150$, a sequência de testes ML produziu resultados consideravelmente superiores aos do método de validação cruzada. Para $T = 500$ a performance de ambos foi comparável, com ligeira superioridade para o que usa testes ML.

Todas as estratégias encontraram problemas para especificar corretamente árvores simuladas com todas as divisões suaves. Uma divisão muito suave seguida de uma rígida aumento o número de especificações errôneas; veja C.1 para detalhes. Mesmo assim, o algoritmo STR-Tree especificado por testes ML superou os outros dois competidores.

Embora a decisão de gerar divisões extremamente suaves no primeiro nó tenha dificultado a especificação da árvore correta, mesmo assim o algoritmo STR-Tree saiu-se relativamente bem em grandes amostras. O grande problema ocorrido durante a aplicação deste algoritmo foi que este não pode identificar a arquitetura correta, assim como as variáveis de divisão, quando simulado uma divisão rígida após uma divisão suave.

Sempre que o algoritmo CART foi submetido a especificação de árvores suaves, a tendência foi criar uma quantidade de nós terminais menor do que a esperada ou mesmo não crescer a árvore. Na situação contrária, mesmo o procedimento de podagem a posteriori não evitou o superajuste.

A estratégia de usar um experimento de validação cruzada do tipo 10-dobras durante a especificação produziu, aparentemente, resultados no algoritmo STR-Tree semelhantes ao CART. Embora a superestimação do número de folhas não tenha sido tão dramática como no CART, a tendência, principalmente em amostras pequenas, foi produzir árvores maiores do que o esperado.

Em grandes amostras e divisões rígidas, a validação cruzada e testes ML produziram resultados semelhantes, muito embora deva ser ressaltado que o custo computacional do primeiro é consideravelmente maior.

5.6 Aplicação à Dados Reais

Nesta seção são apresentadas aplicações para alguns conjuntos de dados reais, alguns bastante utilizados na literatura de análise de regressão. Uma breve descrição dos conjuntos utilizados é dada abaixo:

– Boston Housing – Valores(medianos) de moradias em 506 distritos da cidade de Boston. Este conjunto é o mesmo utilizado em [23] para explicar a árvore de regressão obtida através do algoritmo CART.

– Cpus – O conjunto de dados Cpus é discutido em [92]. O propósito de aplicar a metodologia de árvores de regressão a este conjunto é criar um modelo capaz de explicar o desempenho de 209 diferentes tipos de Cpus de acordo com diferentes características do *hardware*.

– Cars – Este conjunto de dados foi retirado da biblioteca MASS presente no pacote R e descreve o preço e outras 25 variáveis mensuradas em 93 modelos de carros novos para o ano de 1993 nos Estados Unidos.

– Auto imports – Este conjunto de dados foi retirado do livro automotivo anual da Ward´s, edição de 1985, e consiste de 195 preços de carros acompanhados de algumas características tais como: consumo de gasolina, largura, comprimento, tamanho do motor, dentre outras. As informações deste conjunto são similares às do conjunto anterior, porém existe uma quantidade maior de variáveis contínuas para inclusão no modelo.

– Abalone – Este é um conjunto de dados biológicos cujo objetivo de aplicar um modelo de regressão é realizar a predição da idade de um abalone a partir de um conjunto de características físicas. Neste conjunto há 4177 casos e sete variáveis preditoras disponíveis no repositório UCI.

Pela escolha dos conjuntos descritos acima, foram consideradas diferentes situações que variam de pequenas amostras à grandes amostras e, em alguns casos, as variáveis preditoras são altamente correlacionadas, o que traz uma dificuldade adicional na seleção das variáveis de transição.

É importante lembrar que, em todas as situações, considerou-se apenas variáveis contínuas como candidatas para divisão dos nós.

O Conjunto de dados *Boston housing* consiste de 506 observações de uma variável resposta univariada (valor mediano de moradias em dólares) e um conjunto de 13 variáveis explanatórias mensuradas em distritos de Boston. Algumas importantes variáveis são: taxa de criminalidade (CRIM), percentual de terra pala lotes (ZN), taxa de impostos (TAX), número médio de quartos (RM), razão professor/aluno (P/T) e percentual da população de baixo status (LSTAT). O interesse inicial na modelagem destes dados pode ser visto em [42], que pretendia verificar o efeito da poluição, medida pela concentração de (NOX), na variável resposta. Os autores ajustaram uma regressão linear que necessitou de transformações nas variáveis respostas, aumentando a dificuldade na interpretação dos resultados. Material sobre Análise de Regressão aplicada a este conjunto pode ser encontrado em [10]

O ajuste da regressão pelo CART a estes dados resultou em um modelo com 9 folhas e 4 profundidades, após um procedimento de podagem a posteriori.

A especificação modelo STR-Tree a este conjunto de dados foi feita usando o ciclo iterativo descrito na seção 5.2. Para encontrar o tamanho adequado da árvore adotou-se o procedimento de redução do tamanho do teste à medida em que novos nós fossem acrescentados. O ajuste final é apresentado na Figura 5.6.

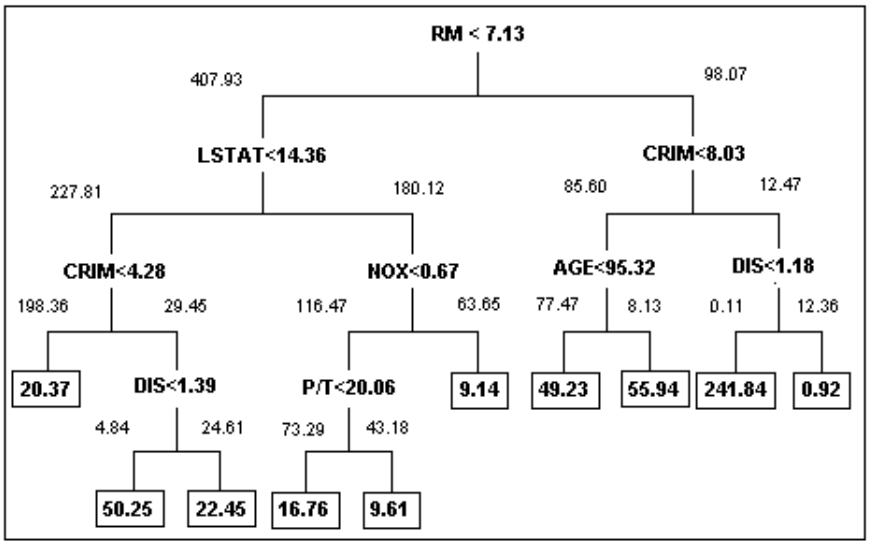

Figura 5.6: Modelo STR-Tree para o Conjunto *Boston Housing*

Por trás da estrutura de árvore mostrada na Figura 5.6 há uma especificação para o modelo MRSTR (Multiple Regime Smooth Transition Regression). Assim, o ciclo de modelagem proposto por [85] pode ser iniciado a partir desta especificação, muito embora não haja garantias de que a interpretabilidade da árvore irá se manter. Entretanto, uma solução intermediária que concilia interpretabilidade e optimalidade na estimação dos parâmetros é realizar uma busca restrita no espaço dos parâmetros em uma vizinhança em torno do vetor de parâmetros estimados.

Os parâmetros não-lineares associados com a árvore na Figura 5.6 são apresentados na Tabela 5.7. Nota-se na Tabela 5.7 que as divisões são mais suaves nos parâmetros no topo da árvore se comparadas com as feitas nas maiores profundidades.

Na Tabela 5.7, os índices dos parâmetros ajudam a localizá-los na saída gráfica da árvore. Assim, c_0 é o parâmetro de locação na raiz(nó 0) que produziu filhores associados aos parâmetrosc_1 and c_2.

Uma diferença importante em relação ao ajuste feito pelo CART é que o modelo STR-Tree pode ajustar um valor para a constante dentro da folha que não pertence ao domínio observado dos dados. De fato, neste conjunto *Boston Housing* esta característica torna-se providencial, pois, no conjunto de dados, o valor máximo das moradias aparece restrito a 50 mil dólares, o que não representa uma situação real.

Para avaliar a incerteza sobre os parâmetros estimados dentro das folhas, utilizou-se a seguinte estratégia: montar intervalos de confiança condicionais ao conhecimento dos parâmetros não-lineares. Deste modo, o erro padrão das

Tabela 5.7: Estimativas dos Parâmetros Não-Lineares do Modelo STR-Tree Ajustado ao Conjunto *Boston Housing*

Parâmetros Não-Lineares	
Locação	Suavidade
$\hat{c}_0 = 7.13_{(0.003)}$ [5]	$\hat{\gamma}_0 = 1.62$
$\hat{c}_1 = 14.36_{(1.166)}$	$\hat{\gamma}_1 = 1.93$
$\hat{c}_2 = 8.03_{(1.085)}$	$\hat{\gamma}_2 = 3.53$
$\hat{c}_3 = 4.28_{(1.086)}$	$\hat{\gamma}_3 = 65.13$
$\hat{c}_4 = 0.67_{(4\times10^{-4})}$	$\hat{\gamma}_4 = 10.63$
$\hat{c}_5 = 95.32_{(4.317)}$	$\hat{\gamma}_5 = 115.36$
$\hat{c}_6 = 1.18_{(0.001)}$	$\hat{\gamma}_6 = 103.41$
$\hat{c}_9 = 20.06_{(5\times10^{-4})}$	$\hat{\gamma}_8 = 245.24$
$\hat{c}_8 = 1.39_{(0.145)}$	$\hat{\gamma}_9 = 5.04$

estimativas das constantes dentro dos nós foi obtido conforme

$$\hat{\sigma}^2 \times \left(Z(\hat{\boldsymbol{\theta}})' Z(\hat{\boldsymbol{\theta}})\right)^{-1} \tag{5-25}$$

onde $\hat{\sigma}^2$ é a estimativa da variância do erro e $Z((\hat{\boldsymbol{\theta}}))$ é uma matriz com o número de colunas igual ao número de folhas e o número de linhas igual ao tamanho da amostra. Cada linha de $Z(.)$, traz a pertinência estimada das observações aos nós terminais.

A soma dos graus de pertinência dentro das folhas é um importante instrumento para identificar candidados a valores atípicos (*outliers*). Um bom exemplo pode ser retirado da Figura 5.6, onde o menor valor para esta estatística (0.11) está associado ao nó 10 que irá conter a regra (nebulosa) de predição na Figura 5.7.

Se		
Condição		
1	Número médio de quartos (RM) é maior que 7.13	&
2	Taxa de criminalidade (CRIM) é maior que 8.03	&
3	Distância para o centro comercial (DIS) é menor que 1.18	
Então		
Valor predito da moradia é	241.84	

Figura 5.7: Regra(nebulosa) de predição associada as observações no interior do nó 10

Os valores medianos dos distritos que estão mais próximos de satisfazer as condições acima são completamente diferentes de quaisquer outros pontos observados em região vizinha. De fato, o conjunto de dados não contém uma observação que siga exatamente a regra descrita na Figura 5.7, mas mesmo assim

podem ser feitas predições, sob uma enorme variabilidade, para valores nesta região do espaço das variáveis.

Para os demais conjuntos, não serão explicitados os modelos obtidos, mas na sequência apresenta-se os resultados encontrados após submetê-los aos diferentes algoritmos de especificação.

Para obter uma figura honesta do desempenho alcançado pelos algoritmos de especificação, foi conduzido um experimento fora-da-amostra que consistiu em retirar 10% da amostra, aplicar os algoritmos, e repetir este procedimento 10 vezes. Isto resultou em um total de 100 medidas do EQM avaliado fora-da-amostra.

A Tabela 5.8 reporta a mediana, o Desvio Absoluto Mediano (DAM), o máximo e o mínimo dos erros quadráticos nestas 100 observações.

Tabela 5.8: Erro Quadrático Médio Fora-da-Amostra dos algoritmos CART e STR-Tree baseados em 100 Observações. observations.

	CART			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	19.29	6.00	9.61	59.56
Cpus	4.15×10^3	2.16×10^3	471.68	4.85×10^4
Car Sales	39.33	21.39	8.50	266.50
Auto Imports	7.75	2.53	2.39	19.55
Abalone	5.67	0.43	4.08	7.61
	STR-Tree/LM			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	14.51	4.25	7.00	50.43
Cpus	2.38×10^3	1.33×10^3	257.92	1.92×10^4
Car Sales	25.71	13.48	3.45	175.44
Auto Imports	9.17	2.11	4.37	27.07
Abalone	5.32	0.51	3.99	6.81
	STR-Tree/CV			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	12.06	2.96	6.49	43.32
Cpus	3.05×10^3	1.94×10^3	280.00	2.67×10^4
Car Sales	26.40	15.68	3.08	169.66
Auto Imports	11.27	3.05	3.94	33.32
Abalone	6.26	0.63	4.21	8.38

Com a exceção do Auto Imports, o STR-Tree model comportou-se melhor do que o CART. O STR-Tree especificado pela sequência de testes ML superou os resultados da especificação por validação cruzada em 4 dos 5 conjuntos.

Entretanto, se o número de folhas no modelo final for usado como medida de custo-complexidade, no mesmo espírito proposto em [23], o modelo STR-Tree/ML é mais parcimonioso do que o CART em 3 dos 5 conjuntos, como pode ser visto na Tabela 5.9.

O modelo STR-Tree/CV gerou árvores menores do que o STR-Tree/LM em 4 dos 5 casos.

A Tabela 5.9 mostra a mediana, DAM, mínimo e máximo do número de nós terminais nos 100 casos analisados.

Tabela 5.9: Número de Folhas Especificados pelo CART e STR-Tree baseados em 100 Observações.

	CART			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	7	1	4	15
Cpus	5	1	2	11
Car Sales	3	0	1	4
Auto Imports	9	2	3	16
Abalone	11	1.5	7	16
	STR-Tree/LM			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	9	1	4	12
Cpus	7	1	4	10
Car Sales	2	0	2	4
Auto Imports	4	0	4	7
Abalone	8	1	4	12
	STR-Tree/CV			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	7	1	4	12
Cpus	3	0	3	9
Car Sales	2	0	2	3
Auto Imports	3	1	2	6
Abalone	2	0	2	10

A Tabela 5.10 mostra a mediana, DAM, mínimo e máximo do tempo computacional, em segundos, gastos na especificação dos 3 modelos propostos, avaliados em 100 observações.

Todos os programas foram executados no Matlab 6.5.1. No caso do CART foi utilizada uma função customizada chamada *treefit* da caixa de ferramentas estatísticas *Statistical Toolbox*. Todo o trabalho computacional foi realizado em um micro Pentium IV, 2.8 GHz com 1 Gb de memória RAM. Pode ser observado através da inspeção da Tabela 5.10 que o gasto computacional com o uso do modelo STR-Tree/CV é consideravelmente maior do que a estratégia utilizada pelo STR-Tree/LM que demonstra ser uma alternativa competitiva ao CART.

Tabela 5.10: Tempo (em segundos) gasto pelos algoritmos CART e STR-Tree com base em 100 observações.

	CART			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	29.69	0.94	22.85	40.43
Cpus	7.31	0.22	6.65	11.97
Car Sales	5.44	0.39	4.61	40.81
Auto Imports	11.02	0.58	8.07	12.36
Abalone	61.72	1.27	42.11	68.45

	STR-Tree/LM			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	38.73	9.48	6.78	145.61
Cpus	28.80	5.06	17.17	66.56
Car Sales	10.93	2.30	1.23	43.52
Auto Imports	26.63	9.49	7.36	65.95
Abalone	91.06	15.56	64.13	495.61

	STR-Tree/CV			
Conjuntos de Dados	Mediana	DAM	Min.	Max.
Boston	1.07×10^3	161	570.00	1.85×10^3
Cpus	197.00	19.9	161.00	604.50
Car Sales	121.00	8.2	92.80	227.10
Auto Imports	393.30	92.2	231.90	824.50
Abalone	645.30	33.9	566.50	3.1202×10^3

5.7 Conclusões

Neste Capítulo foi proposto um novo modelo que combina aspectos do CART (Classification and Regression Trees) e STR (Smooth Transition Regression). O Modelo foi denominado STR-Tree e sua idéia principal é a substituição da função indicadora utilizada pela metodologia CART pela função logística. O modelo resultante pode ser analisado como uma regressão com transição suave entre múltiplos regimes.

Uma detalhada análise das propriedades assintóticas dos estimadores foi apresentada e o procedimento para construção do modelo, baseado em uma seqüência de testes do tipo Multiplicadores de Lagrange (ML), foi desenvolvido. Uma forma alternativa de especificação baseada em um experimento de validação cruzada também foi discutida e um experimento de Monte Carlo foi conduzido para avaliar o desempenho das metodologias propostas, comparando-as com o CART.

O modelo STR-Tree superou o CART quando foi considerada a seleção correta da arquitetura de árvores simuladas. Além do mais, o teste ML aparenta ser uma alternativa promissora a usual especificação por validação cruzada.

Adicionalmente, o algoritmo de estimação teve comportamento satisfatório em pequenas amostras.

Ao utilizar dados reais para testar o desempenho do modelo STR-Tree este mostrou habilidade preditiva superior ao CART. Finalmente, aponta-se o modelo STR-Tree como uma ferramenta a ser utilizada nas Florestas Aleatórias (*Random Forests*) desenvolvidas recentemente em [21].