AVALIAÇÃO DE MÉTODOS NÃO PARAMÉTRICOS PARA ESTIMAÇÃO DE MODELOS ADITIVOS

Marco Aurelio Valles Leal, Tainá Aparecida Zalourensi, Willian Luís de Oliveira (Orientador), e-mail: wloliveira@uem.br, Carlos Aparecido dos Santos (Co-Orientador), e-mail: casantos@uem.br

Universidade Estadual de Maringá / Centro de Ciências Exatas / Maringá, PR.

Probabilidade e Estatística – Inferência não paramétrica

Palavras-chave: regressão, modelo aditivo, suavizadores

Resumo:

É comum, nas mais diversas áreas, investigar e modelar a relação entre variáveis. O modelo mais simples é denominado modelo de regressão linear simples, entretanto nem sempre a relação existente é perfeitamente linear. Neste contexto, é possível flexibilizar o modelo de regressão linear modelando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico, sendo esta nova classe de modelos denominada modelos aditivos. Desta forma, este projeto visa introduzir os modelos aditivos, apresentando algumas técnicas de suavização utilizadas para ajustar modelos no contexto não paramétrico, em particular os modelos aditivos, assim como suas principais características e aplicações. Por fim, com o objetivo de validar a metodologia estudada, será realizado um estudo de simulação.

Introdução

Análise de regressão é uma técnica estatística para investigar e modelar a relação entre variáveis, sendo essa uma técnica amplamente utilizada na estatística. Usualmente, é de interesse apenas uma variável, chamada de variável resposta ou dependente, e desejamos estudar como esta variável depende de um conjunto de variáveis observáveis, chamadas de variáveis explicativas ou independentes. Nesse contexto, os modelos de regressão linear simples e múltipla podem ser utilizados.

Nota-se, porém, que em muitos casos a relação existente entre a variável resposta (média) e cada uma das variáveis explicativas não é perfeitamente linear e determinar uma função que estima a relação correta existente nem sempre é fácil. Uma alternativa é flexibilizar o modelo de regressão linear, modelando a dependência da variável resposta com cada uma das variáveis explicativas em um contexto não paramétrico. Esta nova classe de modelos

é dita modelos aditivos e mantêm a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos.
As funções suaves do componente sistemático do modelo podem ser estimadas através de um suavizador (*smoother*). Entretanto, algumas técnicas de suavização podem não ser viáveis em alguns problemas.
Portanto, o objetivo do presente trabalho é apresentar os modelos aditivos e estudar as principais técnicas de suavização existentes utilizadas para ajustar modelos no contexto não paramétrico, em particular os modelos aditivos, apresentando suas principais características e aplicações. O *software* estatístico R (*R Team Core*) será utilizado para a realização de todas as análises do estudo.

Materiais e métodos

Uma das mais populares e úteis ferramentas em análise de dados é a análise de regressão, cujo objetivo é investigar e modelar a relação entre variáveis (características). No caso mais simples, são consideradas duas variáveis, a variável resposta (dependente) Y e a variável preditora (independente) X, e o objetivo é descrever a dependência da média de Y como função de X, considerando *n* observações dessas variáveis. Para isto, assumimos que a média de Y é uma função linear de X.
O modelo que é aplicado na estrutura de regressão mais simples é denominado por modelo de regressão linear simples (MONTGOMERY, et. al., 2012). Já o modelo de regressão linear que envolve mais de uma variável explicativa (X_1, X_2, ..., X_p) é chamado de modelo de regressão linear múltipla e considera que a variável resposta Y depende das variáveis preditoras (explicativas) X_1, X_2, ..., X_p, através da relação linear

$$Y = b_0 + b_1X_1 + b_2X_2 +...+ b_pX_p + \varepsilon,$$

em que (b_0, b_1, b_2,..., b_p) são constantes desconhecidas que devem ser estimadas, considerando as *n* observações das variáveis do estudo e ε é o erro aleatório do modelo, normalmente distribuído com média 0 e variância constante. Entretanto, em algumas aplicações a dependência da esperança de Y em X_1, X_2, ..., X_p é de longe linear, não sendo adequado o uso do modelo linear. Poderíamos adicionar alguma outra relação, mas geralmente é difícil encontrar a forma mais apropriada. Neste contexto, utilizamos os modelos aditivos, uma extensão do modelo de regressão linear e que pertence a uma classe mais geral denominada modelos aditivos generalizados (HASTIE & TIBSHIRANI, 1990).
A ideia principal é substituir a função linear usual da covariável com algum tipo específico de função suavizadora, consistindo assim na soma de tais funções suavizadoras, ou seja, temos a relação

$$Y = b_0 + f_1(X_1) + f_2(X_2) +...+ f_p(X_p) + \varepsilon,$$

em que $f_i(X_i)$ são funções desconhecidas. Se trata de um modelo não paramétrico no sentido de não impor parâmetros para as funções, mas sim estimá-los de forma iterativa com uso de suavizadores do gráfico de dispersão.

Os suavizadores do gráfico de dispersão podem ser usados para descrição, ajudando a encontrar uma tendência no comportamento do gráfico de dispersão de Y versus X (dependência funcional sem impor presunções paramétricas). Consideramos neste trabalho três suavizadores: a técnica regressão polinomial local, conhecida como *loess* (CLEVELAND, 1979), a suavização com núcleos (suavizadores tipo *kernel*) e os *splines* de regressão (*regression splines*) com destaque para os *splines* cúbicos.
Quando mais de uma covariável está disponível para predizer a resposta, utilizamos o algoritmo retroajuste (*backfitting*), utilizado em modelos aditivos (BUJA et. al., 1989; HASTIE & TIBSHIRANI,1990), para estimar cada função suave em um cenário não paramétrico, além do intercepto.

Resultados e Discussão

Verificamos o desempenho dos suavizadores estudados, em alguns cenários simulados. Em um desses cenários, foram geradas 100 observações para cada uma das variáveis independentes X_1 e X_2, no intervalo [0, π], ou seja, uma sequência de valores entre 0 e π. A variável resposta Y foi definida como

$$y = 10 + \cos(x_1) + \text{sen}(x_2) + \varepsilon,$$

sendo ε o erro, com média 0 e variância constante igual a 1. É importante notar que y foi definido como uma soma de funções das covariáveis, ou seja, tem-se um modelo aditivo.
Usando o algoritmo de retroajuste estudado, tendo como base o suavizador *Loess* e o suavizador *Kernel*, e considerando a convergência quando a diferença entre as curvas ajustadas no passo anterior e no passo atual, é inferior a 0,01%, obtemos as seguintes curvas ajustadas:

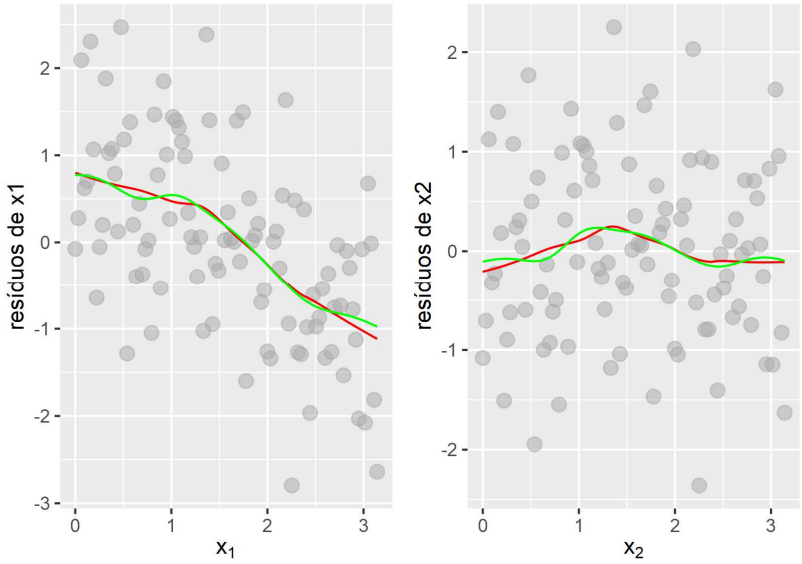

Figura 1 – Gráfico de dispersão dos resíduos parciais de X_1 e X_2, respectivamente e as curvas das funções estimadas pelo método Loess (vermelho) e Kernel (verde)

A partir do gráfico acima é possível notar que o ajuste não foi exatamente o mesmo entre as técnicas, porém ambos os suavizadores representam bem a tendência dos dados. A curva verde, obtida pelo método *Kernel*, tem uma variabilidade maior, enquanto que a vermelha, obtida pelo *Loess* está mais suave. Dessa forma, para os dados gerados e técnicas apresentadas, o ajuste através do método *Loess* foi mais satisfatório, pela curva ter uma oscilação menor, para este cenário estudado.

Conclusões

Para investigar e modelar relações entre variáveis o modelo de regressão linear pode ser utilizado, porém quando essa relação não possui forma linear, uma alternativa é o uso de ferramentas que não impõem suposições paramétricas. Nesse contexto, existem técnicas de suavização que podem ser utilizadas, inclusive na estimação das funções do componente sistemático dos modelos aditivos. Caso haja mais de uma covariável para predizer Y, o algoritmo de retroajuste pode ser uma solução.
Para validar a metodologia estudada, um estudo de simulação foi realizado, apresentando o ajuste de funções para um modelo aditivo através do algoritmo de retroajuste com base em duas das técnicas de suavização estudadas: *Kernel* e *Loess*, sendo que o último apresentou, visualmente, um melhor ajuste. Vale ressaltar que as análises realizadas em relação à qualidade do ajuste foram estritamente gráficas. Existem medidas numéricas adequadas para fazer essa análise, que serão abordadas futuramente.

Agradecimentos

Agradecemos o Departamento de Estatística pelo suporte neste trabalho.

Referências

BUJA, A., HASTIE, T. & TIBSHIRANI, R. (1989). **Linear smoothers and additive models**. The Annals of Statistics, 17, 453-510.

CLEVELAND, W. S. (1979). **Robust locally weighted regression and smoothing scatterplots**. Journal of the American Statistical Association, 74, 829-836.

HASTIE, T. J. & TIBSHIRANI, R. J. (1990). **Generalized additive models**, volume 43. Chapman and Hall, Ltd., London. ISBN 0-412-34390-8.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G**. Introduction to Linear Regression Analysis**. 5th Edition. John Wiley & Sons, 2012.
TEAM, R. CORE. R: **A language and environment for statistical computing**. (2013).