**Introdução**

**Slide 1**

**Tópico 1 - Contexto:** A análise de regressão é amplamente utilizada na estatística que visa explorar e modelar a relação entre variáveis. Em particular, o modelo de regressão linear múltipla é aplicado nas mais diversas áreas do conhecimento como, por exemplo, nas ciências biológicas, físicas e químicas, na economia e na engenharia.

**Tópico 1 - Definição:** A análise de regressão tem como objetivo descrever a rela¸ c˜ ao entre uma variável resposta de interesse (y) e um conjunto de p variáveis preditoras. A variável resposta está relacionada às covariáveis na seguinte relação linear:

**Tópico 1 - Comentário sobre o modelo de regressão:** onde os parâmetros $\beta\_0, \beta\_1,\ldots,\beta\_p$ são os coeficientes de regressão e o termo $\varepsilon$ é um erro aleatório normalmente distribuído com média zero e variância $\sigma^2$ constante e serem não correlacionados. O parâmetro $\beta\_j$ representa a mudança esperada na resposta $y$ a cada mudança unitária de $x\_j$ quando todas as outras variáveis regressoras são constantes. os coeficientes de regressão são parâmetros desconhecidos e podem ser estimados utilizando, dentre outras técnicas, o método dos mínimos quadrados.

**Tópico 2 – Definição:** Como a verdadeira função que descreve o relacionamento entre *y* e as *p* covariáveisé desconhecida, mas em certos intervalos das variáveis regressoras, o modelo de regressão linear é uma aproximação adequada para a verdadeira função desconhecida.

**Tópico 3 – Motivação:** Uma solução seria acrescentar uma transformação nas variáveis regressoras, por exemplo, o método de transformação de Box-Cox. Contudo, determinar uma transformação que represente a correta relação existente nem sempre é uma tarefa fácil. Outra possibilidade 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.

**Tópico 4 – Introdução ao Modelo Aditivo:** o componente sistemático é formado por uma soma de funções suaves não especificadas das covariáveis, além de um termo aleatório. Esta nova classe de modelos é dita modelos aditivos e mantém a característica dos modelos de regressão lineares de serem aditivos nos efeitos preditivos.

**Tópico 4 – Definição:** onde os erros $\varepsilon$ são independentes com média zero e variância constante $\sigma^2$. Cada $f\_j(x\_j)$ é uma função univariada arbitrária.

**Tópico 5/6 – Comentário Geral**: Mas não precisa ter necessariamente um comportamento linear. De certa forma, os modelos aditivos podem ser vistos como uma flexibilização do modelo de regressão linear.

**Objetivos**

**Slide 2**

As funções $f\_j$ devem ser estimadas por meio de suavizadores, que estimam uma tendência menos variável e descreve sua dependência em relação à variável resposta. Algumas das técnicas mais conhecidas para obter as estimativas suavizadas são: suavizador \*bin\*, média móvel, linha móvel, suavizador, \*Loess\* ou \*Lowess\*, de \*Kernel\* e \*Splines\* de regressão. De fato, a "suavização" de cada técnica supracitada depende da escolha (estimativa) do seu respectivo "parâmetro de suavização". Em algumas técnicas, por exemplo, dependendo do valor do parâmetro de suavização, pode-se ter uma reta estimada ou até uma interpolação dos dados. Em termos gerais, pode-se dizer que a variância e o viés na estimação/predição do modelo suavizado dependem do parâmetro suavizador. Este

problema é, de certa forma, análogo à questão de quantas variáveis preditoras colocar em uma equação de regressão.

Logo, torna-se de suma importância, além do estudo comparativo das técnicas de suavização no ajuste em modelos aditivos no contexto não paramétrico, também a estimação do parâmetro suavizador de cada técnica. Adicionalmente, o estudo do desempenho na predição dessas técnicas também é pertinente partindo da hipótese prévia que nem sempre o melhor ajuste resulta em melhor poder preditivo.

Dito isto, descreve-se a seguir os objetivos de estudo deste trabalho nesta direção.

**Metodologia - Suavizadores**

**Slide 3**

**Tópico 5 – Usos do Suavizador**:

Descrição - Um suavizador em diagrama de dispersão pode ser usado para melhorar a aparência visual do gráfico de $x$ \*versus\* $y$ para encontrar uma tendência nos dados;

Estimação – Estimar a dependência da esperança de $y$.

**Slide 4**

**Topicos – Explicação Geral Suavizador Categórico:** Pode não parecer que simplesmente realizar as médias seja um processo de suavização, mas este conceito é a base para a configuração mais geral, já que a maioria dos suavizadores tenta "imitar" a média da categoria por meio da média local, ou seja, realizar a média dos valores de $y$, tendo os valores preditores próximos dos valores alvo. Esta média é feita nas vizinhanças em torno do valor alvo.

**Tópico 5 – Tamanho da Vizinhança:** Intuitivamente, grandes vizinhanças produzirão estimativas com variância pequena mas potencialmente com um grande viés e inversamente quando adotado vizinhanças pequenas. Portanto, temos uma troca fundamental entre variância e viés estipulada pelo parâmetro suavizador.

**Tópico 6 – Como realizar a média na vizinhança:** pois os suavizadores diferem principalmente pelo jeito de realizar as médias. Algumas das técnicas mais conhecidas para obter as estimativas suavizadas são: suavizador \*bin\*, média móvel, linha móvel, suavizador \*Loess\* ou \*Lowess\*, suavizadores de \*Kernel\* e \*Splines\* de regressão.

**Slide 4**

**Tópico 1 – Explicação Geral Suavizador Bin:** , também conhecido como regressograma, imita um suavizador categórico, particionando os valores preditores em regiões disjuntas e, então, realizando a média da resposta em cada região.

**Tópico 2 – Explicação Geral Média Móvel:** é outra técnica que leva em conta o cálculo da média. É muito comum utilizar uma vizinhança/região de $(2k + 1)$ observações, $k$ para a esquerda e $k$ para a direita de cada observação, no qual o valor de $k$ tem um comportamento de troca entre suavidade e qualidade do ajuste. Um problema comum encontrado na média móvel é o viés. Uma saída é usar pesos para dar mais importância às vizinhanças mais próximas.

**Tópico 3 – Explicação Geral Linha Móvel:** na qual novamente são definidas as vizinhanças para cada ponto, tipicamente os $k$ pontos mais próximos de cada lado. Nesse caso é mais interessante considerar a proporção de pontos em cada vizinhança, ou seja, $w = \dfrac{(2k+1)}{n}$, denominado \textit{span}. Então, ajusta-se uma linha de regressão aos pontos de cada região, que é usada para encontrar o valor predito suavizado para o ponto de interesse.

**Metodologia – Suavizador Loess**

**Slide 5**

Destaca-se que nessa técnica deve-se ter atenção à escolha do valor do \textit{span}. Um valor muito pequeno faz com que a curva seja muito irregular e tenha variância alta. Por outro lado, um valor muito grande fará com que a curva seja sobre-suavizada, podendo não se ajustar bem aos dados e resultando em perda de informações e viés alto.

**Metodologia – Suavizador Kernel**

**Slide 6**

Tópico 1 - O suavizador \*kernel\* representa a sequência de pesos descrevendo a forma da função peso com o auxílio de uma função densidade; com um parâmetro de escala que ajusta o tamanho e a forma dos pesos perto de $x\_0$.

Tópico 2 - onde $b$ é o tamanho da vizinhança (parâmetro -- de escala -- suavizador), e $K$ uma função kernel, ou seja, uma função densidade. Existem diferentes escolhas para $K$, geralmente usa-se a densidade de uma Normal, tendo-se, assim, um \*kernel\* Gaussiano.

**Metodologia – Suavizador Splines de Regressão**

**Slide 7**

**Tópico 2 -** Dessa forma, é possível modelar com polinômios mais simples as curvas mais complexas. Os \textit{splines} dependem, principalmente, do grau do polinômio e do número e localização dos nós.

Essa técnica é interessante, pois, tem uma maior flexibilidade para o ajuste dos modelos em comparação com o modelo de regressão polinomial ou linear e, após a determinação da localização e quantidade de nós, o modelo é de fácil ajuste.

**Tópico 4 -** contínuo e contendo primeira e segunda derivadas contínuas nos nós. As \textit{splines} cúbicas são as de menor ordem nas quais a descontinuidade nos nós são suficientemente suaves para não serem vistas a olho nu, então, a não ser que seja necessárias mais derivadas suavizadas, existe pouca justificativa para utilizar \textit{splines} de maior ordem.

**Tópico 5** - sendo mais importante o número de nós do que sua localização. Salienta-se que incluir mais nós que o necessário pode resultar em uma piora do ajuste do modelo. Existem algumas maneiras para fazer essas escolhas, como, por exemplo, colocar os nós nos quantis das variável preditora (três nós interiores nos três quartis).

**Metodologia – Seleção de Parametros – Função de Risco**

**Slide 7**

**Tópico 3 -** ou seja, aquele que possui o menor risco. Um modelo pode interpolar os dados e, mesmo assim, possuir um baixo poder preditivo (IZBICKI E SANTOS, 2020).

**Tópico 4** - Nesse sentido, usa-se o critério do risco quadrático para averiguar a qualidade da função. Assim, escolhe-se uma função $g$ em uma classe de candidatos $G$ que tenha um bom poder preditivo (baixo risco quadrático). Dessa maneira, visa-se evitar modelos que tenham sub ou super-ajuste.

**Tópico 5** - onde $g$ é escolhida, a fim de minimizar o EQM acima, sendo $\hat{Y}\_i = g(X\_i)$ o valor predito de $Y\_i$ por $g$. Nota-se que a predição é feita para cada observação, após o ajuste do modelo utilizando todos os dados disponíveis. Contudo, este estimador, se empregado para realizar a seleção de modelos pode levar um super-ajuste (ajuste perfeito aos dados).

**Metodologia – Seleção de Parametros – Validação Cruzada**

**Slide 8**

Tópico 1 - Utiliza-se os dados de treinamento para estimar a regressão e se avalia o erro quadrático médio por meio do conjunto de validação. Este procedimento de divisão é chamado de \*data splitting\* (IZBICKI E SANTOS, 2020).

Tópico 2 - Consiste em dividir a base dados em K conjuntos disjuntos, realizando uma varredura nos conjuntos. Treina-se o modelo com C = K-1 conjuntos e se valida com o conjunto que ficou de fora. Deve-se realizar o rodízio dos K conjuntos até que todos os dados sejam vistos como dados de treino e validação.

Tópico 3 - , no qual o modelo é ajustado utilizando todas as observações com exceção da i-ésima delas, sendo um caso particular da técnica anterior \*k-fold\* no caso de $K=n-1$.

Tópico 4 – Procedimento para seleção do parâmetro suavizador:

1.1.1 - Considere para os dados de treino todas as observações, exceto i-ésima delas, consequentemente, ter-se-á apenas uma observação compondo os dados de validação.

1.2 - Ao final, um vetor de tamanho $n$ das diferenças quadráticas. O erro quadrático médio, para o respectivo parâmetro suavizador, será a média:

2 - O melhor parâmetro suavizador $p$ será aquele que gerou o menor EQM, dentre todos os cadidatos possíveis em seu domínio.

**Metodologia – Seleção das técnicas de suavização**

**Slide 8**

Tópico 1 - que fornecerá o melhor ajuste para um determinado suavizador. Em seguida, consideraremos duas métricas para selecionar a melhor técnica de suavização.

Tópico 2 - Destacar-se-á a melhor técnica de suavização aquela obter o menor $EQM\_{c}$ dentre as técnicas comparadas.

Tópico 3 - Novamente, a melhor técnica de suavização será aquela que obter o menor $EQM\_{loocv}$ dentre as técnicas comparadas.

De fato, o menor $EQM\_{c}$ apresentará a técnica com melhor ajuste aos dados e o menor $EQM\_{loocv}$ escolherá aquela de maior poder preditivo.

**Resultados e Discussão – Estudo de Simulação**

**Slide 9**

Tópico 3 - Sendo assim, um valor que indica a proporção de observações a serem utilizadas próximas, ou ainda, nos arredores do ponto de estimação de interesse. Para os \*splines\* de regressão, o parâmetro de suavização indicado, normalmente, pela quantidade de nós utilizados para os ajustes. No presente trabalho, a quantidade de nós fora considerado, sendo que a posição e localização deste nós fora mantidas fixas, equidistantes espaçadas, em K percentis distintos.

Tópico 4 - avaliando, visualmente, os comportamentos das curvas em diagramas de dispersão.

Tópico 5 - encontrar-se-á um parâmetro de suavização que forneça a ocorrência do menor erro quadrático médio ($EQM\_{loocv}$) possível, desta forma, evitando um super-ajuste do modelo. Ademais, ajustes serão executados tendo em consideração tais parâmetros e, em seguida, calcular-se-á os erros quadráticos médios ($EQM\_c$) para cada técnica suavizadora, entre os valores observados e estimativas do modelo.

Tópico 6 - contabilizando a quantidade de vezes em que cada técnica apresenta o menor erro quadrático médio. Por exemplo, para o primeiro cenário, será gerado mil amostras aleatórias de tamanho $n$. Para cada amostra será empregado o procedimento acima, salvando seus repectivos erros quadráticos médio. Ao final da simulação, será contabilizado se a ocorrência do erro quadrático médio em cada técnica foi mínima e, por fim, comparar e verificar qual técnica obtém o melhor resultado em uma simulação de mil amostras.

Tópico 7/8 - Considerar-se-a a técnica com o melhor desempenho de predição a que obtiver o menor $EQM\_{loocv}$. E o melhor ajuste a técnicas que possuir o menor $EQM\_{loocv}$ para, assim, concluir qual dos suavizadores são mais aderentes aos dados.

Tópico 9 - Não obstante, serão gerados nove sub-cenários, valendo-se da combinação de três tamanhos amostrais (150, 250 e 350), em três valores de desvio padrão distintos.

**Considerações Finais**

**Slide 9**

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.

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.

As técnicas de suavização \*kernel\*, \*loess\* e splines de regressão, em particular, os de grau um e grau três foram apresentadas, que são utilizadas para estimar as funções presentes em modelos aditivos. Fora introduzido um método para obtenção no melhor parâmetro suavizador, a fim de evitar sub ou super-ajuste, realizado por meio de método de validação cruzada. Duas métricas foram abordadas: uma para verificar a qualidade de predição, erro quadrático médio obtidos por validação cruzada, \*leave one out cross validation\* ($EQM\_{loocv}$) e a outra para apurar a qualidade dos ajustes, erro quadrático médio completo ($EQM\_c$).

Para validar a metodologia estudada, foram realizadas análises em dados simulados e dados reais. Em dados simulados de diferentes cenários, foram observados os resultados em relação ao comportamento de técnicas de suavização, comparando os modelos obtidos, por meio das métricas introduzidas anteriormente. Ainda, verificou-se que o ajuste mais adequado para descrever o comportamento dos dados não obtém necessariamente o melhor poder preditivo.

Por meio dos resultados obtidos do estudo de simulação, pode-se concluir que, ao avaliar e comparar os ajustes considerando a métrica de qualidade de predição ($EQM\_{loocv}$) e a métrica de qualidade de ajuste ($EQM\_c$), tanto para o Cenário 1, quanto para o Cenário 2, o suavizador que obteve o melhor poder preditivo foi o \*splines\* de regressão cúbico. Levando em consideração o método com melhor qualidade de ajustes, os suavizadores com \*kernel\* se destacou em ambos os cenários.

Finalmente, os métodos discutidos foram aplicados em dados reais, Aplicação 1 e Aplicação 2, nos quais, mais uma vez, os suavizadores foram avaliados. Validou-se qual apresenta o melhor poder preditivo e qual representa de forma mais adequada os dados. Ressalta-se que para a Aplicação 1, a técnica que obteve o melhor poder preditvo e o melhor ajuste fora o \*splines\* de regressão cúbico. Para a Aplicação 2, o suavizador que se destacou por obter o melhor poder preditivo (menor $EQM\_{loocv}$ entre os suavizadores) foi \*splines\* de regressão linear. Em contrapartida, o que denotou melhor qualidade de ajuste (menor $EQM\_c$ dentre os suavizadores) foi o método \*kernel\* gaussiano.

Ademais, outras métricas para validação da qualidade de predição e adequabilidade dos modelos podem ser adotadas. Existem outras técnicas que podem ser adotadas para seleção dos parâmetros de suavização que não foram discutidas neste trabalho. Outrossim, especificamente para os suavizadores \*splines\* de regressão, além da seleção da quantidade de nós, a localização dos nós (que foram mantidas fixas e equidistantes nos k percentis possíveis) pode ser avaliada a fim de obter um melhor ajuste. Aliás, todas discussões realizadas podem ser extendidas, quando mais de uma covariável está disponível para predizer a resposta. Frequentemente, utiliza-se o algoritmo de retroajuste (\*backfitting\*, HASTIE & TIBSHIRANI, 1990) para estimar cada função suave $f\_j$ em um cenário não paramétrico.