



MODELOS LINEARES GENERALIZADOS: UMA APLICAÇÃO A DADOS VOCAIS

A by Amauri Neto

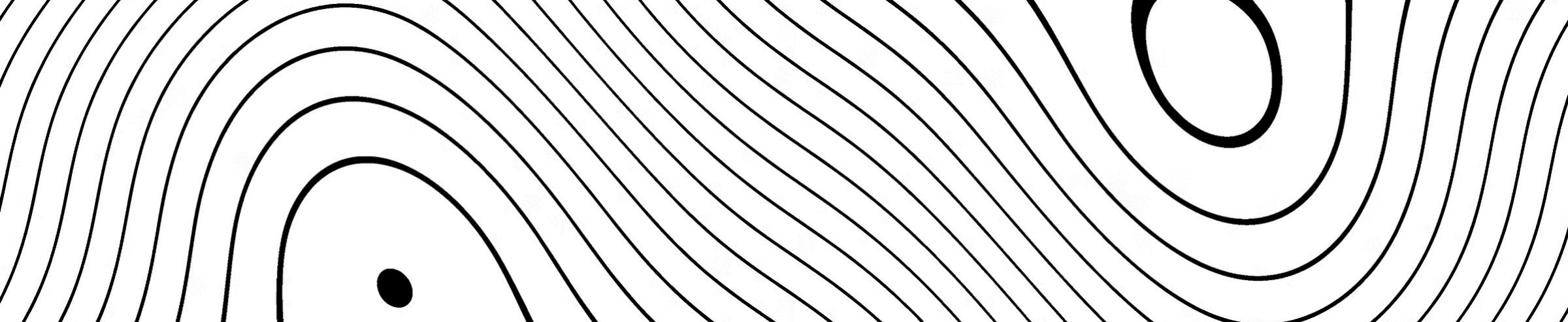
A voz humana é uma ferramenta poderosa de comunicação, expressão e identidade pessoal. No entanto, a produção vocal pode ser afetada por diversas condições que resultam em irregularidades na voz.

Essas irregularidades podem se manifestar de várias formas como:

- rouquidão
- aspereza
- tremor
- variação no tom ou volume

Principais causas:

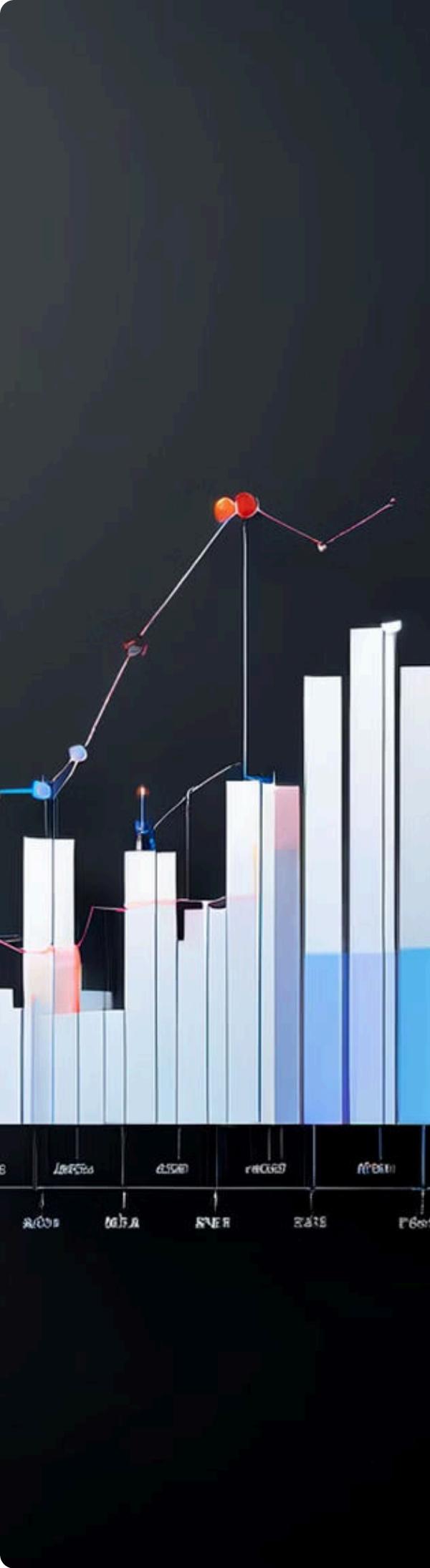
- abusos vocais
- infecções respiratórias
- lesões nas cordas vocais
- paralisia laríngea
- condições neurológicas



Compreender a origem e o impacto dessas irregularidades é fundamental para o diagnóstico e tratamento adequado, visando a recuperação ou melhoria da qualidade vocal e, consequentemente, da comunicação e qualidade de vida do indivíduo.

Metodologia: Modelos Lineares Generalizados (MLG)

Modelos Lineares Generalizados (MLG) são uma extensão dos modelos lineares e foram introduzidos por John Nelder e Robert Wedderburn em 1972 de modo que permitem uma maior flexibilidade na modelagem de variáveis dependentes que não seguem uma distribuição normal mas que pertencem a família exponencial



1

Função de Ligação

A função de ligação $g(\mu)$ relaciona a média da variável resposta μ com a combinação linear dos preditores η .

2

Distribuição da Família Exponencial

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

onde θ é o parâmetro canônico, ϕ é o parâmetro de dispersão, e $b(\theta)$ e $c(y, \phi)$ são funções específicas da distribuição.

3

Estrutura Linear e Componente de Variância

Nos MLGs, a estrutura linear é representada por $\eta = X\beta$. A variância da variável resposta é uma função da média μ e do parâmetro de dispersão ϕ :

$$\text{Var}(Y) = \phi V(\mu)$$

onde $V(\mu)$ é a função de variância. No caso da distribuição gama, temos $V(\mu) = \mu^2$.

Criação e análise do modelo:

1

Métodos de Estimação de Parâmetros

Os parâmetros dos MLGs são tipicamente estimados usando o método da máxima verossimilhança, que, comparado aos mínimos quadrados, a máxima verossimilhança é mais flexível para distribuições não normais.

2

Função Desvio (Deviance)

A função desvio (deviance) mede a qualidade do ajuste do modelo comparando um modelo ajustado com um modelo saturado.

$$2 \sum_{i=1}^n \left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right)$$

3

Avaliação de Ajuste do Modelo

Para avaliar a adequação do modelo, usamos critérios como o AIC (Akaike Information Criterion) e o BIC (Bayesian Information Criterion), além de testes de bondade de ajuste e a análise do desvio.

4

Overfitting e Underfitting

Overfitting: ocorre quando o modelo se ajusta excessivamente aos dados de treinamento.

Underfitting: ocorre quando o modelo é muito simples.



Interpretabilidade dos Resultados

Tornar os resultados dos MLGs interpretáveis envolve a escolha apropriada de funções de ligação e transformações, além de técnicas de visualização de dados.

Modelo Linear Generalizado e Modelo Linear Comum

MLGs permitem a modelagem de variáveis resposta que seguem distribuições não normais e utilizam funções de ligação para transformar relações não lineares em lineares.

Modelagem

Flexibilidade em modelar diferentes tipos de dados e variáveis resposta.

Variedade de funções de ligação

A escolha da função de ligação depende da natureza da variável resposta e da distribuição dos dados.

Desta forma será aplicado o modelo linear generalizado para dados de 131 pacientes com irregularidades na voz

A base de dados se encontra disponível no kaggle.

Variáveis

- **jitter:** Nervosismo do paciente
- **shimmer:** mede a oscilação de atingir tons afinados
- **GNE:** impacto genético na voz do paciente
- **Irregularity:** variável resposta que mede a irregularidade da voz do paciente
- **Noise:** Barulho emitido pela voz do paciente
- **OverallSeverity:** Gravidade geral da voz do paciente

Resumo das variáveis:

Variável	Tipo	Mínimo	Média	Máximo
Jitter	contínua	0,03	1,961	31,58
Shimmer	contínua	2,7	13,821	62,42
GNE	contínua	0,15	0,5905	0,96
Irregularity	contínua	0,5	1,34	2,79
Noise	contínua	0,16	0,9224	2,66
OverallSeverity	contínua	0,47	1,206	2,6

Gráfico de densidade



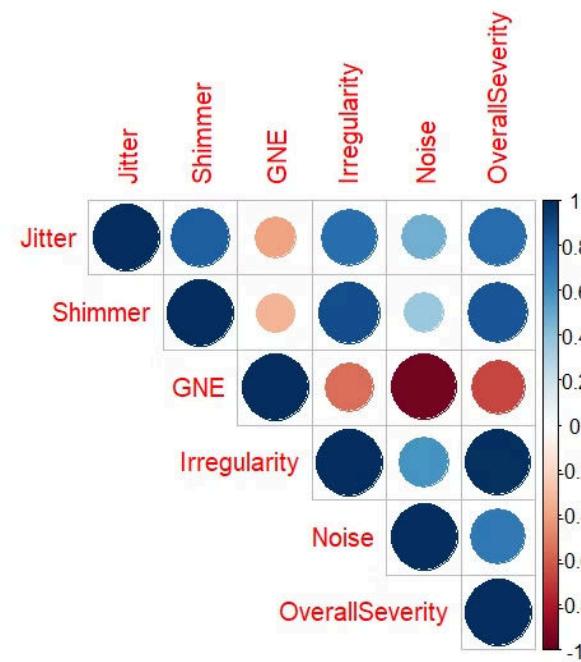
Teste de Normalidade

O teste de Shapiro-Wilk é um teste estatístico utilizado para avaliar a normalidade de uma amostra de dados. Ele verifica se a distribuição dos dados se desvia de uma distribuição normal.

- Hipótese nula (H_0): Os dados seguem uma distribuição normal.
- Hipótese alternativa (H_1): Os dados não seguem uma distribuição normal.

Aplicando o teste à Variável Irregularity, temos um p-valor de 1.71e-07 que não nos dá uma probabilidade suficiente para aceitar a H_0 , então aceitamos H_1 e seguimos com a análise.

Correlações entre as variáveis



**Desta forma, foi a variável
Irregularity foi modelada a partir da
função Gama e função de ligação
log.**

Testando todas as possíveis combinações de variáveis para o modelo, temos que só 3 foram estatisticamente significativas .

Modelo ajustado com as variáveis significativas

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.79376	0.020838	-38.098	< 2e-16 ***
Jitter	-0.016900	0.001789	-9.446	2.30e-16 ***
Noise	-0.118851	0.012839	-9.257	6.62e-16 ***
OverallSeverity	0.968587	0.022800	42.481	< 2e-16 ***

Com um AIC de -276.76 e Deviance de 0.54245 com 127 Graus de Liberdade.

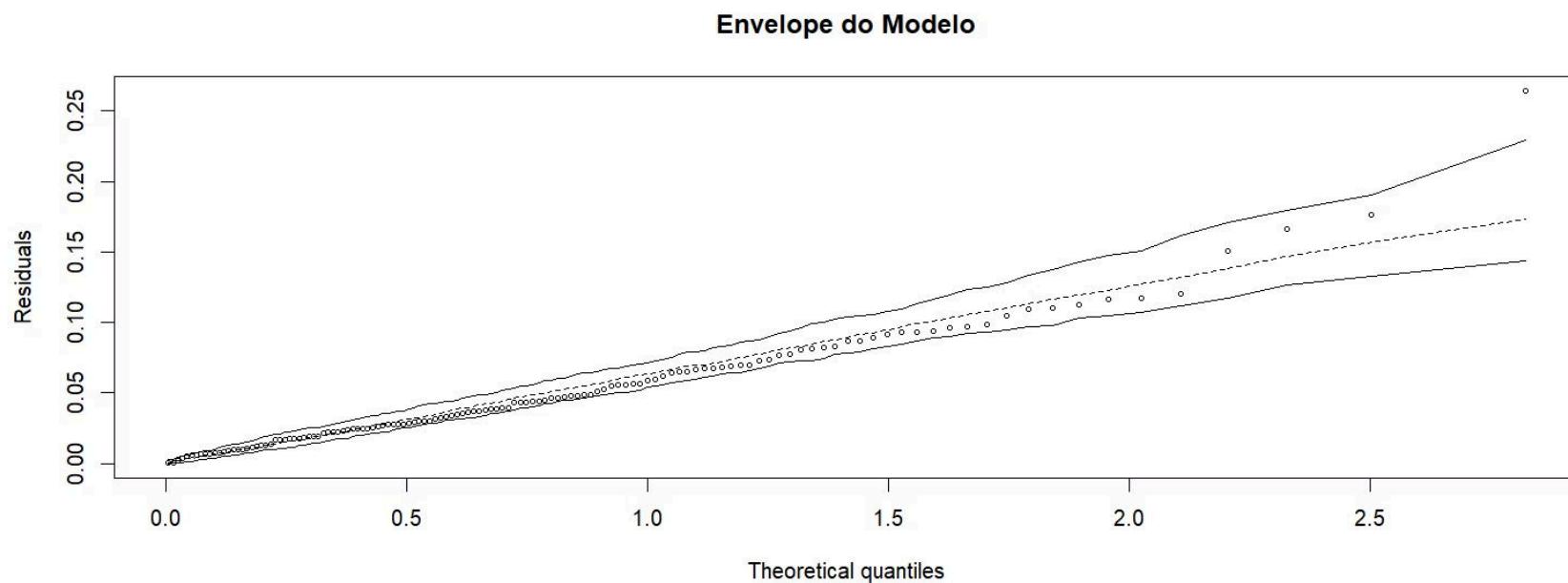
Testes de multicolaridade

Valores dos fatores de inflação

	Jitter	Noise	OverallSeverity
VIF	2.354515	1.976416	3.573493

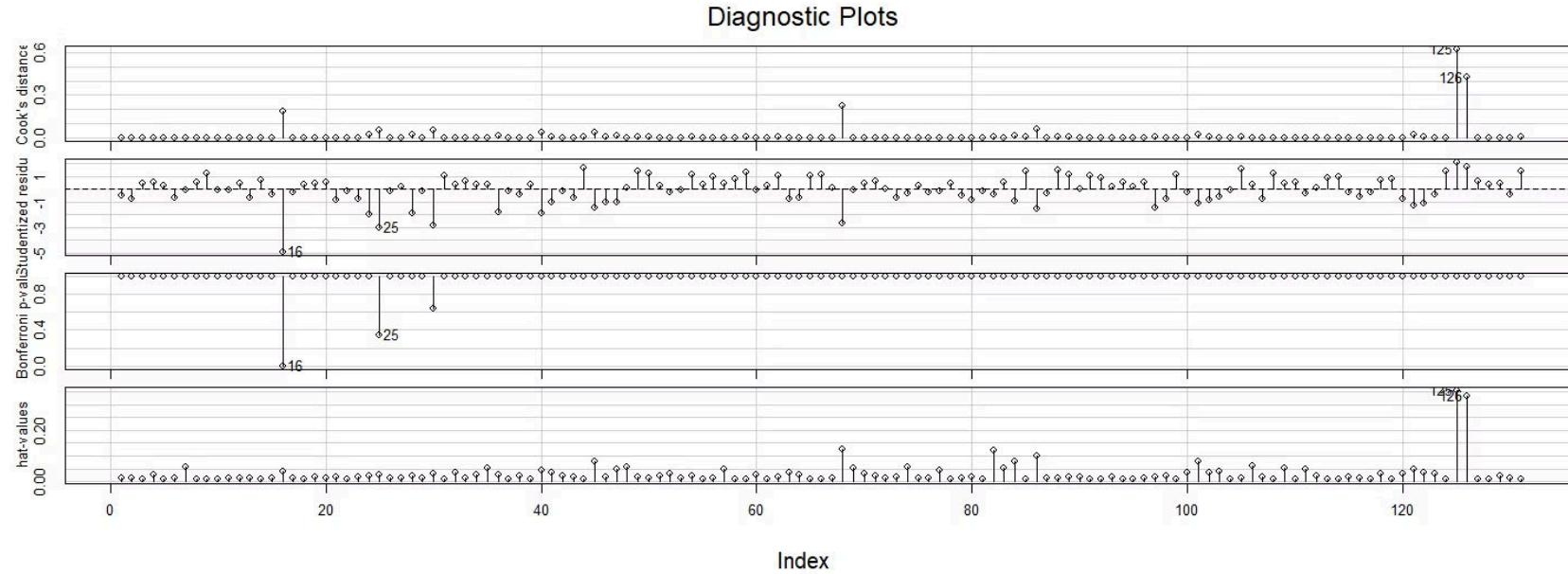
Como nenhuma das variáveis tem um VIF maior que 10, não há evidencia de multicolinearidade.

Gráfico de envelope:



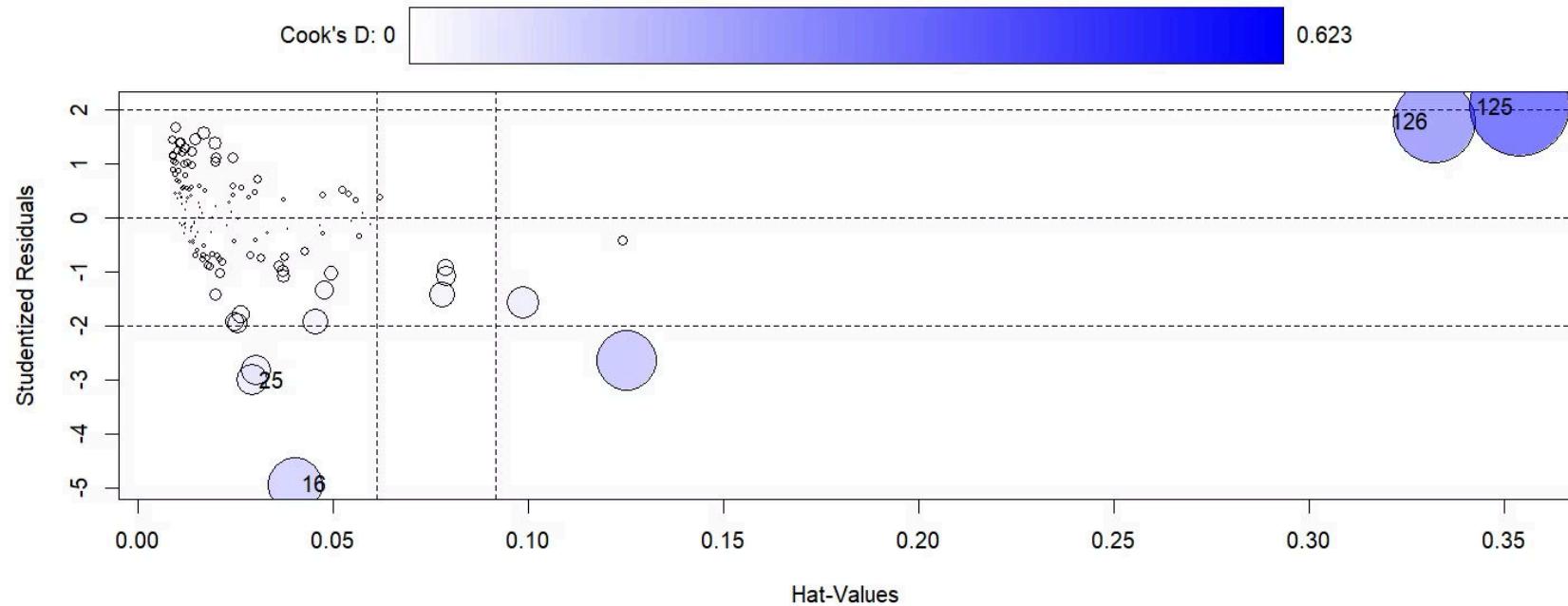
O gráfico de envelope nos dá a ideia de um bom ajuste no modelo, não contendo apenas 1 valor que representa 0,76% dos valores que é inferior ao limite de 5%.

Diagnóstico do modelo:



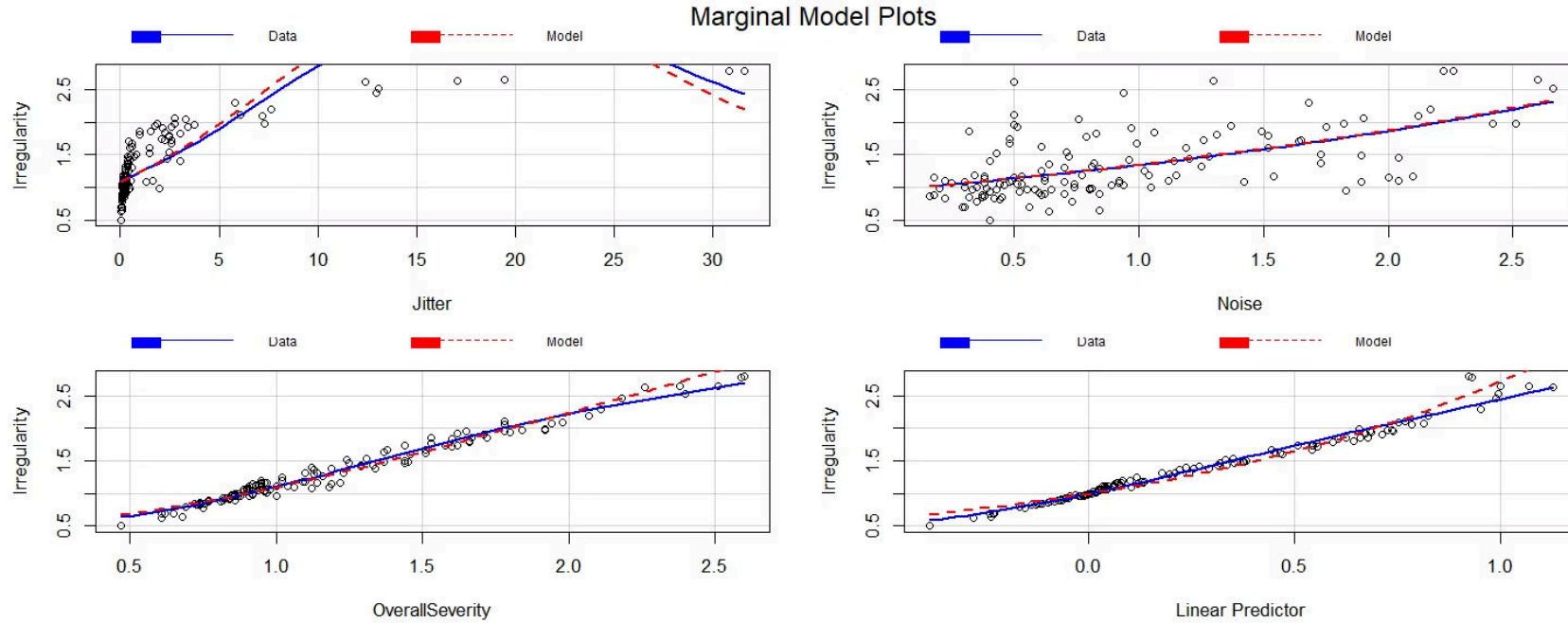
Nos gráficos de diagnóstico observa-se que o ponto 26 é um possível outlier, e os pontos 125 e 126 são os com maiores valores de distância de Cook e hat-values, ou seja, apresenta os maiores valores nas variáveis independentes.

Gráfico de influência:



Observando o gráfico de influência acima vemos que como já esperado, os pontos 125 e 126 são os que mais influenciam no modelo.

Gráficos Marginais:



Os gráficos acima destacam a relação entre a variável dependente e cada uma das independentes assim como o preditor linear em que podemos observar uma boa relação entre os dados reais e as estimativas do modelo.

Interpretação dos estimadores

Para podermos interpretar os parâmetros do modelo, é necessário retorná-los para sua escala de origem, ou seja, aplicar a função inversa da função de ligação.

Valores dos parâmetros	(Intercept)	Jitter	Noise	OverallSeverity
$\exp(\beta)$	0.452	0.9832	0.8880	2.6342

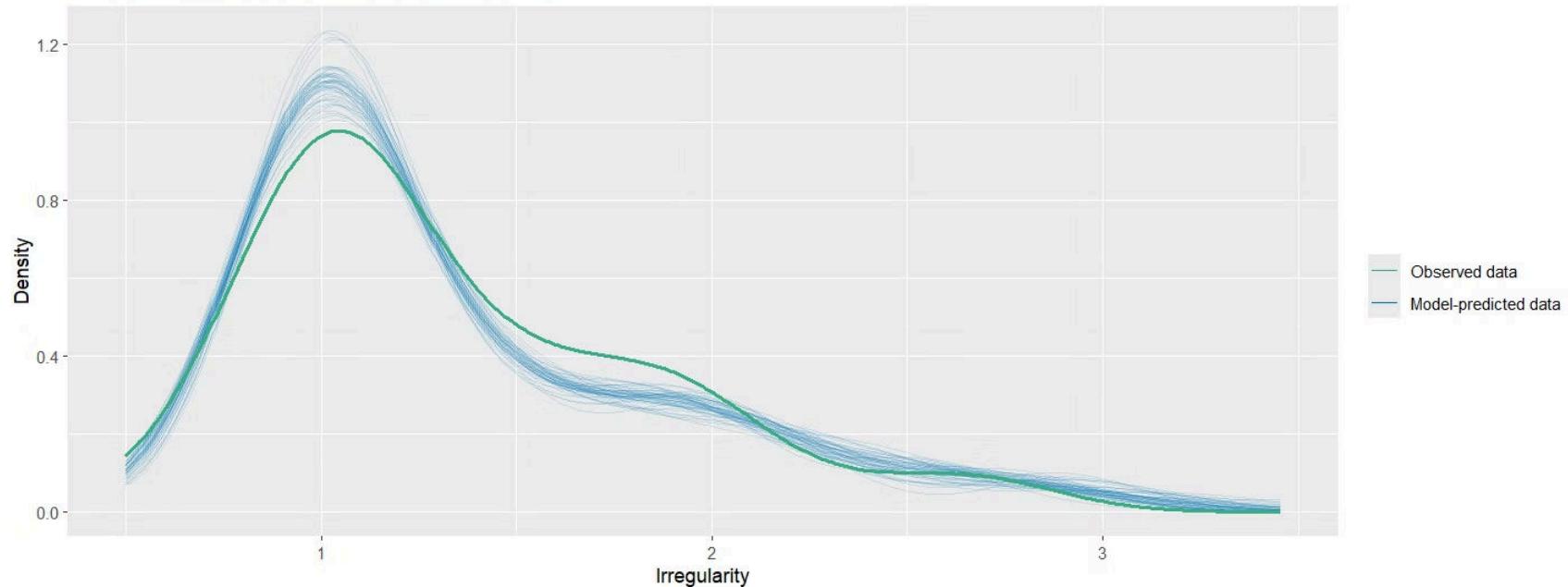
Neste sentido, a cada unidade de nervosismo do paciente(jitter) sua irregularidade na voz aumenta cerca de 0.9832, para o barulho(Noise) aumenta 0.888 e para cada unidade de gravidade na voz, sua irregularidade vocal aumenta 2.6342. Caso todas as variáveis fossem 0 sua irregularidade vocal seria de 0.452(intercept).

Conclusão:

De modo Geral, o modelo mostrou uma boa adequação aos dados, modelando de forma aceitável a variável resposta(Irregularity) e apresentando bons resultados no diagnóstico, entretanto é importante salientar o ponto 26 que pode ser um possível outlier diminuindo assim a eficiência do modelo.

Posterior Predictive Check

Model-predicted lines should resemble observed data line



O gráfico acima mostra 100 simulações (em azul) utilizando o modelo criado e a densidade dos dados(em verde) para uma melhor visualização do comportamento preditivo e adequação do modelo.



FIM!