Universidade Federal do Paraná

Andressa Luiza Cordeiro

Estudos de caso com aplicações de diferentes modelos de regressão para dados com resposta no intervalo unitário

Curitiba
2025

Andressa Luiza Cordeiro

Estudos de caso com aplicações de diferentes modelos de regressão para dados com resposta no intervalo unitário

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Cesar Augusto Taconeli

Curitiba
2025

Agradecimentos

A Deus, por ter nos concedido vida, oportunidade e capacidade para passar por mais esta etapa; sem ele isto não seria possível.

Às nossas famílias, pelo apoio e paciência no decorrer desses anos de graduação, assim como em toda a nossa vida.

Ao Professor Doutor Cesar Augusto Taconeli, pela paciência, disponibilidade de seu tempo e compartilhamento de seu conhecimento acadêmico.

Ao Professor Doutor Pedro Henrique Toledo de Oliveira Sousa, pela disponibilidade em participar da banca deste trabalho.

"É justo que muito custe o que muito vale."
(Santa Teresa de Jesus)

Resumo

Neste trabalho, foram analisadas duas bases de dados distintas: uma contendo dados de felicidade em diversos países e outra composta por dados sintéticos relacionados à satisfação de funcionários. Em ambas as bases, a variável resposta é contínua e restrita ao intervalo (0,1), o que justificou a utilização da abordagem GAMLSS (*Generalized Additive Models for Location, Scale and Shape*), combinada a modelos para respostas com essa característica. Esse método permite modelar não apenas o parâmetro de locação, mas também os parâmetros de escala e forma em função das covariáveis, proporcionando uma modelagem mais flexível. No desenvolvimento da análise, foi seguido um protocolo de modelagem que teve início com a comparação de diferentes distribuições candidatas: Beta, Beta Generalizada Tipo 1 (GB1), Beta Generalizada Tipo 2 (GB2), Logito-Normal e Simplex. Após a escolha da distribuição mais adequada, foram aplicados suavizadores nas covariáveis utilizadas na modelagem da média, visando capturar possíveis efeitos não lineares. Em seguida, realizou-se a avaliação da inclusão de variáveis nos demais parâmetros a fim de aprimorar o ajuste. A escolha final do modelo foi baseada no Critério de Informação de Akaike (AIC), enquanto a qualidade do ajuste foi avaliada por meio da análise dos resíduos e, principalmente, dos gráficos do tipo *worm plot*. A seleção de modelos em ambas as bases apontou a Beta Generalizada Tipo 1 (GB1) como a distribuição mais adequada. No entanto, cada base exigiu ajustes específicos na modelagem de seus parâmetros. No caso da base de felicidade, o modelo apresentou um desempenho significativamente mais satisfatório em termos de ajuste, enquanto a base sintética, possivelmente por suas características artificiais, resultou em um ajuste inferior e com maiores inadequações.

Palavras-chave: GAMLSS. Regressão. Resíduos. Intervalo unitário.

Sumário

1 Introdução

Modelos de regressão permitem mensurar a relação entre uma variável de interesse e um conjunto de variáveis explicativas observáveis (MONTGOMERY; PECK; VINING, 2012). Essa relação é formalizada por meio de uma função matemática que combina as variáveis explicativas com o objetivo de quantificar seus efeitos sobre a resposta ou, ainda, prever valores não observados da variável dependente com base em novos valores para as covariáveis (WEBER, 2022).

Dentre os diversos tipos de modelos de regressão utilizados, o modelo de regressão linear descreve a relação linear entre as variáveis (CHARNET et al., 2008). Sua boa aceitação se deve, entre outros fatores, à sua simplicidade para implementação e interpretação dos resultados. Este modelo também se destaca por sua simplicidade computacional. Essas características tornam os modelos de regressão linear uma ferramenta fundamental em Estatística, com grande aplicabilidade em diversas áreas (MONTGOMERY; PECK; VINING, 2012). Neste tipo de regressão, a relação entre a resposta e as variáveis é assumida como linear. Outras suposições que devem ser feitas são a de que, condicional aos valores das variáveis explicativas, a resposta apresenta variância constante, com distribuição normal e de que os erros não sejam correlacionados entre si (GREENE, 2002).

Quando essas suposições não forem atendidas ou quando as limitações afetarem significativamente os resultados, é necessário explorar outras técnicas de modelagem. Afinal, existem modelos apropriados para diferentes tipos de dados e, particularmente, uma grande variedade de distribuições para a resposta. Assim, para contornar os problemas apresentados em modelos lineares, existem os modelos lineares generalizados (GLMs), que são uma classe de modelos estatísticos que possuem uma função de ligação (identidade, logito, logarítmica, etc.) e permitem que a variável dependente tenha uma distribuição pertencente à família exponencial. No caso particular da função de ligação identidade e distribuição normal, temos o modelo de regressão linear. Outros modelos podem ser ajustadas, como a regressão logística, com distribuição binomial e função de ligação logito; a regressão de Poisson, com distribuição de Poisson e função de ligação logarítmica, dentre outros (MCCULLAGH; NELDER, 1989).

Os GLMs (*Generalized Linear Models*) foram formalizados por Nelder e Wedderburn (1972), em um trabalho fundamental para a expansão do uso desses modelos, permitindo realizar modelagens mais gerais e em diversas áreas do conhecimento. Os GLMs são flexíveis o suficiente para acomodar diversos tipos de respostas e são tradicionalmente aplicados a variáveis aleatórias contínuas que representam taxas, proporções ou índices, sendo amplamente utilizados em áreas como ciências sociais, agronomia e psicometria (BONAT; RIBEIRO; ZEVIANI, 2013). No entanto, uma limitação importante desses modelos é a dificuldade em lidar com variáveis cuja distribuição não pertencem à família exponencial.

Uma estratégia frequentemente adotada para contornar essa limitação é transformar a variável resposta de modo que ela assuma valores em todo o conjunto dos números reais. Contudo, tal transformação pode dificultar a interpretação dos parâmetros e aumentar o risco de extrapolações inadequadas nas previsões (MENEZES; FURRIEL, 2019). Por isso, em situações como essa, é mais apropriado utilizar modelos de regressão específicos para o tipo de dado analisado, oferecendo maior flexibilidade e melhor adequação ao comportamento da variável resposta

Uma extensão importante dos GLMs é representada pelos Modelos Aditivos Generalizados (GAMs, *Generalized Additive Model*), propostos por Hastie e Tibshirani (1986). Esses modelos ampliam a estrutura dos GLMs ao permitir que os efeitos das covariáveis sobre a média da resposta sejam representados por funções suavizadas, geralmente estimadas por *splines*. Com isso, os GAMs são capazes de capturar relações não lineares entre as variáveis explicativas e a resposta, mantendo a estrutura aditiva e a interpretação individual dos efeitos. Essa flexibilidade torna os GAMs particularmente úteis em contextos onde não há um conhecimento prévio claro sobre a forma funcional da relação entre as variáveis, permitindo que os dados revelem essas estruturas de forma semi-paramétrica (HASTIE; TIBSHIRANI, 1990).

Ainda assim, tanto os GLMs quanto os GAMs permanecem restritos à modelagem da média da distribuição da variável resposta. Para contextos em que se deseja modelar também outros parâmetros da distribuição como a dispersão, a assimetria ou a curtose, surgem os Modelos Aditivos para Localização, Dispersão e Forma (GAMLSS, *Generalized Additive Model for location, scale and shape*), introduzidos por Rigby e Stasinopoulos (2005). Os GAMLSS permitem especificar distribuições muito mais amplas que a família exponencial, e permitem que cada parâmetro da distribuição seja modelado por meio de funções lineares ou não lineares das covariáveis. Dessa forma, os GAMLSS oferecem uma estrutura altamente flexível e robusta para lidar com dados complexos, como aqueles que apresentam variância heterogênea, assimetria acentuada ou excesso de zeros (STASINOPOULOS et al., 2017). Para variáveis resposta contínuas que assumem valores estritamente no intervalo unitário (0,1), algumas distribuições são sugeridas, tais como Beta, Beta Generalizada do Tipo 1 e 2, Logito-Normal e Simplex.

O modelo de regressão Beta, proposto por Ferrari e Cribari-Neto (2004), assume que a média da variável resposta pode ser relacionada a um conjunto de covariáveis por meio de uma função de ligação apropriada, geralmente a logito, embora outras funções também possam ser utilizadas. Além disso, a regressão Beta permite modelar um parâmetro de precisão, que influencia a variância da resposta. Isso possibilita lidar com heterocedasticidade, uma vez que a variância é função da média e da precisão.

A distribuição Beta Generalizada Tipo 1 possui quatro parâmetros, é uma extensão da distribuição Beta tradicional (que possui apenas dois) e foi introduzidos por Ospina e Ferrari (2010). Uma característica importante é sua capacidade de ajustar distribuições

com maior ou menor concentração de massa de probabilidade em diferentes regiões do suporte, o que a torna útil para modelar proporções com comportamento extremo ou atípico em relação à distribuição Beta padrão.

A distribuição Beta Generalizada Tipo 2 é uma das famílias de distribuições contínuas mais flexíveis. Seu suporte é o intervalo $(0, \infty)$, mas através de reparametrizações pode ser aplicada a diferentes contextos, incluindo proporções. Ela possui quatro parâmetros, o que proporciona grande versatilidade na modelagem de diferentes formas de distribuição, incluindo caudas pesadas e diferentes graus de assimetria. Tem capacidade de ajustar dados com dispersão ainda mais variável e caudas mais extensas que a Beta Generalizada Tipo 1 (MCDONALD; XU, 1995).

A distribuição Logito-Normal é outra alternativa para modelar variáveis resposta contínuas no intervalo (0,1), assumindo que a transformação logito da variável resposta segue uma distribuição normal. Essa abordagem oferece flexibilidade para capturar diferentes níveis de assimetria e heterocedasticidade, sendo útil em casos onde a distribuição Beta ou suas generalizações não fornecem bom ajuste. Segundo Atkinson (1985), a Logito-Normal tem se mostrado eficaz na modelagem de dados de proporções, especialmente quando há concentração de observações próximas dos limites de zero e um, e quando transformações lineares não são suficientes para representar adequadamente a variância.

Outra alternativa importante é a regressão Simplex, proposta originalmente por Barndorff-Nielsen e Jørgensen (1991), que assume que a variável resposta segue uma distribuição Simplex, adequada para dados contínuos em (0,1), especialmente quando a variância apresenta comportamento não monotônico em relação à média. A regressão Simplex tem sido usada como alternativa à regressão Beta em casos de *overfitting*, sensibilidade a *outliers* ou estruturas de dispersão não modeladas adequadamente pela Beta (FERRARI; LEMONTE; CYSNEIROS, 2011).

Objetivo Geral

Comparar o desempenho de diferentes modelos de regressão para dados com resposta no intervalo unitário (0,1) em duas bases de dados, com o propósito de discutir os usos e verificar qual deles melhor se ajusta aos dados.

Objetivos Específicos

a) Pesquisar e apresentar diferentes tipos de modelos de regressão para dados com resposta no intervalo restrito;

b) Utilização da metodologia GAMLSS para o ajuste de modelos de regressão mais gerais para respostas no intervalo (0,1);

c) Aplicar os modelos pesquisados em bases de dados de felicidade mundial e de satisfação de empregados de uma empresa;

d) Comparar os ajustes e identificar o melhor modelo;

e) Discutir a aplicabilidade das modelagens, técnicas utilizadas e suas limitações.

2 Material e Métodos

2.1 Material

Para as análises propostas no projeto, os dois conjuntos de dados utilizados foram extarídos da plataforma Kaggle e os recursos computacionais disponíveis para o desenvolvimento do projeto estão descritos a seguir.

2.1.1 Conjunto de Dados

2.1.1.1 World Happines Report

O *World Happiness Report* é um relatório que avalia o estado da felicidade em escala global publicado pelo Centro de Pesquisa em Bem-Estar da Universidade de Oxford, em parceria com a Gallup (HELLIWELL et al., 2022). Sua primeira edição foi publicada em 2012, sendo lançada em comemoração ao Dia Internacional da Felicidade, celebrado em 20 de março. As classificações e pontuações de felicidade presentes no relatório são construídas a partir de diversos indicadores, sendo a principal fonte de dados o questionário do *Gallup World Poll*. A base dessas pontuações é a resposta à pergunta central de avaliação de vida, conhecida como Escada de Cantril, que convida os entrevistados a imaginarem uma escada com degraus numerados de 0 a 10, onde 0 representa a pior vida possível e 10 a melhor vida possível, e a indicarem em qual degrau sentem que estão atualmente. Após o cálculo da pontuação de felicidade, o relatório analisa quanto cada um dos fatores explicativos contribui para elevar a avaliação da felicidade média de um país em comparação à de uma nação hipotética denominada "Distopia", caracterizada pelos menores valores médios globais para cada variável considerada.

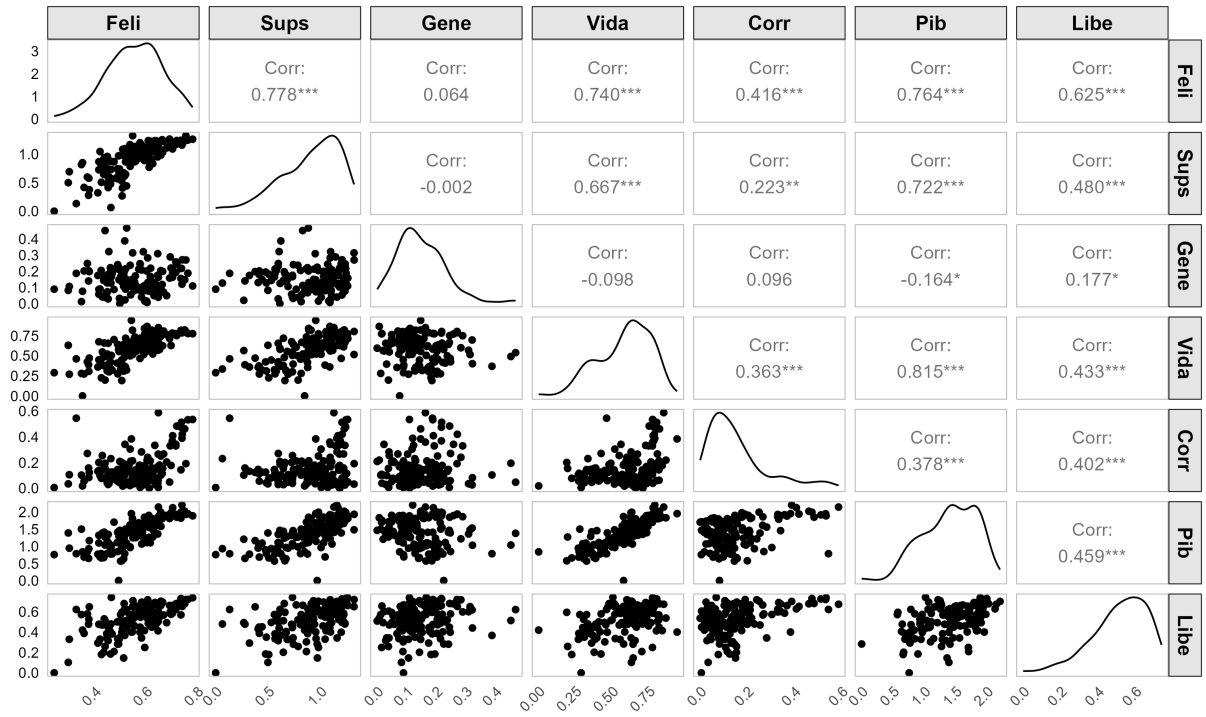
Para este projeto, foi utilizada a base de dados de 2022, que conta com 146 países. A variável resposta, score de felicidade, é aferida numa escala de 0 a 10. Porém, para garantir que as observações respeitem as restrições de valor dos modelos estudados, foi realizada uma transformação linear que ajustou os dados simplesmente dividindo os scores por 10.

As covariáveis utilizadas estão relacionadas na Tabela 1 e o gráfico de correlação na Figura 1.

Tabela 1 – Tabela com descrição das covariáveis (numéricas) para a base de dados de felicidade

COVARIÁVEL	DESCRIÇÃO
Pib	Produto Interno Bruto: Medida econômica que calcula a média da produção econômica por pessoa em um país ou região durante um período específico.
Sups	Apoio Social: Amigos, família e redes de suporte comunitário.
Vida	Expectativa de vida saudável: Número médio de anos que uma pessoa pode esperar viver em plena saúde.
Libe	Liberdade para fazer escolhas: Liberdade que têm para tomar decisões importantes em suas vidas, como escolher onde viver, o que fazer profissionalmente, e como gastar o tempo e os recursos.
Gene	Generosidade: Tendência das pessoas em um país para serem generosas, como a disposição para doar dinheiro ou tempo para ajudar os outros.
Corr	Ausência de Corrupção: Inverso da percepção de corrupção (nível de corrupção existente nas instituições públicas e privadas de seu país).

Figura 1 – Matriz de dispersão e correlação entre as variáveis do World Happiness Report 2022



2.1.1.2 Employee Satisfaction Survey Data

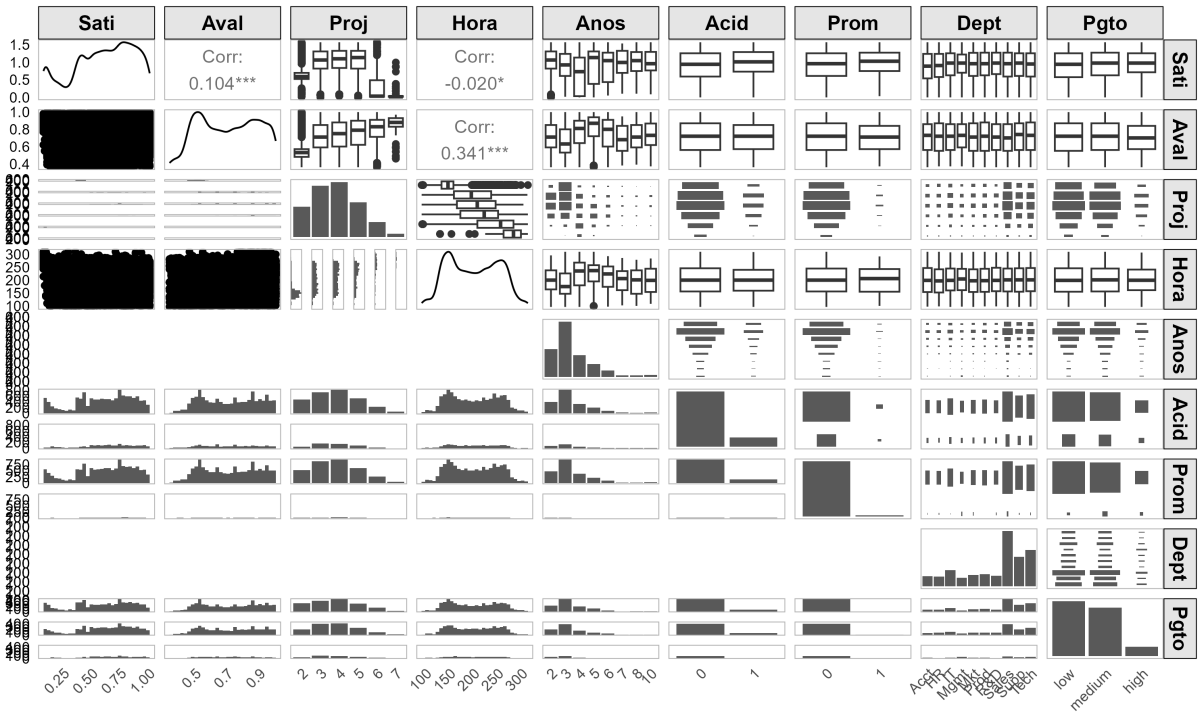
O *Employee Satisfaction Survey Data* é uma pesquisa de satisfação dos funcionários de uma empresa. Ela inclui fatores que influenciam a satisfação dos funcionários e que podem ser usados para compreender vários aspectos do ambiente de trabalho. Este conjunto de dados, embora tenha sido gerado de forma sintética, possui a variável resposta (*satisfaction level*) no intervalo unitário (0,1), o que permite demonstrar as aplicações dos modelos de regressão desejados.

O conjunto de dados possui 15.787 observações, e as covariáveis utilizadas são as observadas na Tabela 2 e o gráfico de correlação na Figura 2.

Tabela 2 – Tabela com descrição das covariáveis (numéricas) para a base de dados de satisfação dos funcionários

COVARIÁVEL	DESCRIÇÃO
Aval	Pontuação mais recente da avaliação de desempenho do funcionário.
Proj	Número de projetos em que o funcionário está atualmente trabalhando.
Hora	Média de horas trabalhadas por mês pelo funcionário.
Anos	Número de anos que o funcionário está na empresa.
Acid	Indica se o funcionário sofreu um acidente de trabalho (1 para sim, 0 para não).
Prom	Indica se o funcionário recebeu uma promoção nos últimos 5 anos (1 para sim, 0 para não).
Dept	Departamento ou a divisão em que o funcionário trabalha.
Pgto	Nível salarial do funcionário (ex.: baixo, médio).

Figura 2 – Matriz de dispersão e correlação entre as variáveis da base de dados de satisfação dos funcionários



2.1.2 Recursos Computacionais

Para as análises estatísticas e tratamento dos dados, foi utilizado o *software* R, versão 4.4.3.

2.2 Métodos

2.2.1 Pré processamento dos dados

Inicialmente, foi realizada a limpeza e seleção das variáveis adequadas para a análise. Foram removidas variáveis consideradas irrelevantes para os objetivos do estudo. Em seguida, as variáveis de interesse foram renomeadas para facilitar a manipulação e a leitura no *software* R. As variáveis mantidas representam aspectos relevantes para explicar a resposta.

2.2.2 Análise exploratória dos dados

Após a seleção das variáveis, foi realizada uma análise exploratória com o objetivo de avaliar as características individuais de cada variável presente na base. Essa etapa buscou identificar a presença de valores ausentes, *outliers*, distribuições assimétricas, e relações não lineares com a variável resposta.

Além disso, foram gerados gráficos e medidas-resumo para investigar o comportamento de cada variável explicativa isoladamente em relação à variável resposta. Essa análise permitiu compreender a estrutura dos dados, avaliar possíveis necessidades de transformação de variáveis e verificar se havia inconsistências que justificassem a substituição ou exclusão de dados.

2.2.3 Modelagem

Dada a flexibilidade e a variedade de caminhos possíveis dentro da modelagem estatística, optou-se por criar um protocolo de modelagem para este trabalho. A definição desse protocolo teve como objetivo garantir um padrão para as bases analisadas, promovendo consistência nas etapas e facilitando a comparação de resultados.

Esse protocolo não é uma regra única ou definitiva, mas sim uma estratégia elaborada especificamente para este estudo, com a intenção de evitar a necessidade de testar um número excessivo de modelos ou decisões arbitrárias ao longo da análise. Com isso, buscamos tornar o processo mais objetivo, organizado e reprodutível.

Como ponto de partida, foram selecionados alguns modelos que se mostram adequados para variáveis resposta contínuas e restritas ao intervalo (0, 1). Os modelos escolhidos para avaliação inicial foram: Beta, Beta Generalizada 1, Beta Generalizada 2, Logito Normal e Simplex. A escolha por essas distribuições se deu devido à sua capacidade de capturar diferentes padrões de assimetria, curtose e heterocedasticidade, características comumente observadas em variáveis deste tipo.

Após o ajuste inicial dos modelos propostos, utilizou-se o critério de informação de Akaike (AIC, *Akaike Criteria Information*) como métrica para seleção do modelo mais adequado. O AIC permite comparar modelos com diferentes complexidades, penalizando o

excesso de parâmetros e favorecendo aqueles que apresentam melhor ajuste com menor sobreajuste.

Com o modelo selecionado, partiu-se para uma análise mais detalhada dos resíduos, utilizando tanto abordagens numéricas quanto visuais. Entre essas abordagens, destaca-se o uso do *wormplot*, um gráfico derivado do gráfico de quantis normalizados, que permite verificar desvios sistemáticos no ajuste do modelo em diferentes regiões da variável resposta. Ele é especialmente útil em modelos GAMLSS, pois facilita a detecção de problemas como assimetria, curtose inadequada ou falhas no ajuste em subgrupos específicos dos dados.

Para a seleção de variáveis, utilizou-se a função `drop1()`, que atua avaliando o impacto da retirada de cada variável do modelo de forma isolada. Essa função testa, uma a uma, as variáveis explicativas presentes no modelo, calculando a mudança no critério de ajuste ao excluir cada uma delas. Assim, é possível identificar quais variáveis têm contribuição significativa e quais poderiam ser removidas sem perda substancial de qualidade no modelo. Esse processo auxilia na simplificação da estrutura do modelo, mantendo apenas os termos relevantes e evitando *overfitting*.

Em seguida, foi realizada a aplicação de suavizadores do tipo `pb()` (*penalized B-splines*) em cada uma das variáveis explicativas e também em algumas combinações delas. A escolha pelo uso desse tipo de suavizador se deve à sua flexibilidade em capturar relações não lineares entre as variáveis explicativas e a resposta, sem necessidade de especificar previamente a forma da relação. O suavizador permite que o modelo se ajuste de maneira mais livre aos dados, ao mesmo tempo em que aplica uma penalização para evitar *overfitting*, mantendo o equilíbrio entre flexibilidade e parcimônia.

Cada modelo com suavizadores foi avaliado individualmente e em combinação, sendo novamente utilizado o critério de informação de Akaike (AIC) para a seleção do modelo final. Esse processo permitiu identificar quais variáveis apresentavam relações não lineares com a variável resposta, bem como determinar a estrutura mais adequada para o modelo final.

Para os demais parâmetros além da média, considerando as particularidades de cada modelo e distribuição, também foi realizada uma seleção das variáveis explicativas. Esse processo utilizou a função `drop1()` e dessa forma foi possível definir, para cada parâmetro (como σ, ν e τ, no caso da distribuição Beta Generalizada 1), quais variáveis deveriam ser mantidas, garantindo um modelo parcimonioso e adequado aos dados.

2.3 Distribuições

2.3.1 Beta

O modelo de regressão beta é baseado na suposição de que a variável resposta, condicional aos valores das explicativas, segue uma distribuição beta. A distribuição

beta (BE) é amplamente reconhecida por sua flexibilidade em modelar proporções, pois sua densidade pode assumir diferentes formas conforme os valores dos parâmetros que caracterizam a distribuição (FERRARI; CRIBARI-NETO, 2004).

A parametrização da distribuição beta é dada por:

$$f(y;\alpha,\beta) = \frac{1}{B(\mu,\sigma)} y^{\mu-1}(1-y)^{\sigma-1}, \qquad 0 < y < 1,$$

onde $\mu > 0, \sigma > 0$.

Sua aplicação no R utiliza a função de ligação `logit` como padrão para `mu.link` e `sigma.link`.

2.3.2 Beta Generalizada Tipo 1

A distribuição beta generalizada tipo 1 (GB1) é uma extensão da distribuição beta tradicional, que oferece maior flexibilidade na modelagem das variáveis, sendo uma boa alternativa quando os dados apresentam características que a beta somente não consegue capturar de forma adequada.

A função densidade de probabilidade da distribuição GB1 é da da por:

$$f(y;\mu,\sigma,\nu,\tau) = \frac{\tau\nu^{\beta} y^{\tau\alpha-1}(1-y^{\tau})^{\beta-1}}{B(\alpha,\beta)\left[\nu+(1-\nu)y^{\tau}\right]^{\alpha+\beta}},$$

onde $0 < y < 1$, $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$, $\tau > 0$ e $\alpha = \frac{\mu(1-\sigma^2)}{\sigma^2}$ e $\beta = \frac{(1-\mu)(1-\sigma^2)}{\sigma^2}$ e $\alpha > 0$, $\beta > 0$.

Sua aplicação no R utiliza a função de ligação logito como padrão para `mu.link` e `sigma.link` e log para `nu.link` e `tau.link`.

2.3.3 Beta Generalizada Tipo 2

A distribuição beta generalizada tipo 2 (GB2) é uma extensão flexível da GB1, definida para variáveis contínuas e positivas ($y > 0$). Também possui os quatro parâmetros e permite uma maior flexibilidade na modelagem de caudas leves ou pesadas, sendo mais apropriada para dados com alta variabilidade ou valores extremos.

A função densidade de probabilidade da distribuição GB2 é dada por:

$$f(y;\mu,\sigma,\nu,\tau) = |\sigma| y^{\sigma\nu-1} \left\{ \mu^{\sigma\nu} B(\nu,\tau) \left[1 + \left(\frac{y}{\mu} \right)^{\sigma} \right]^{\nu+\tau} \right\}^{-1},$$

onde $y > 0$, $\mu > 0$, $\sigma > 0$, $\nu > 0$ e $\tau > 0$.

Sua aplicação no R, assim como a GB1, utiliza a função de ligação identidade como padrão para `mu.link` e `log` para `sigma.link`, `nu.link` e `tau.link`.

2.3.4 Logito-Normal

A distribuição Logito-Normal (LOGITNO) permite controlar a assimetria da distribuição de maneira mais flexível que uma beta padrão e oferece um melhor ajuste quando os dados têm caudas mais pesadas ou formas mais distorcidas. Ela se baseia na suposição de que a transformação logito da variável seja normalmente distribuída.

A função densidade de probabilidade da LOGITNO é dada por:

$$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}y(1-y)} \exp\left(-\frac{1}{2\sigma^2}\left[\log\left(\frac{y}{1-y}\right) - \log\left(\frac{\mu}{1-\mu}\right)\right]^2\right),$$

onde $0 < y < 1$, $0 < \mu < 1$ e $\sigma > 0$.

Sua aplicação no R utiliza função de ligação `logit` como padrão para `mu.link` e log como função de ligação para `sigma.link`.

2.3.5 Simplex

A distribuição Simplex (SIMPLEX) é menos conhecida que a distribuição Beta, mas oferece vantagens específicas quando o interesse é modelar dados com variâncias heterocedásticas uma vez que a variância da distribuição é modelada como uma função da distância quadrática entre os valores observados e a média.

A função densidade de probabilidade da SIMPLEX é dada por:

$$f(y;\mu,\sigma) = \frac{1}{\left(2\pi\sigma^2 y^3(1-y)^3\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\right),$$

onde $0 < y < 1$, $0 < \mu < 1$ e $\sigma > 0$

Sua aplicação no R utiliza a função de ligação `logit` como padrão para `mu.link` e `log` para `sigma.link`.

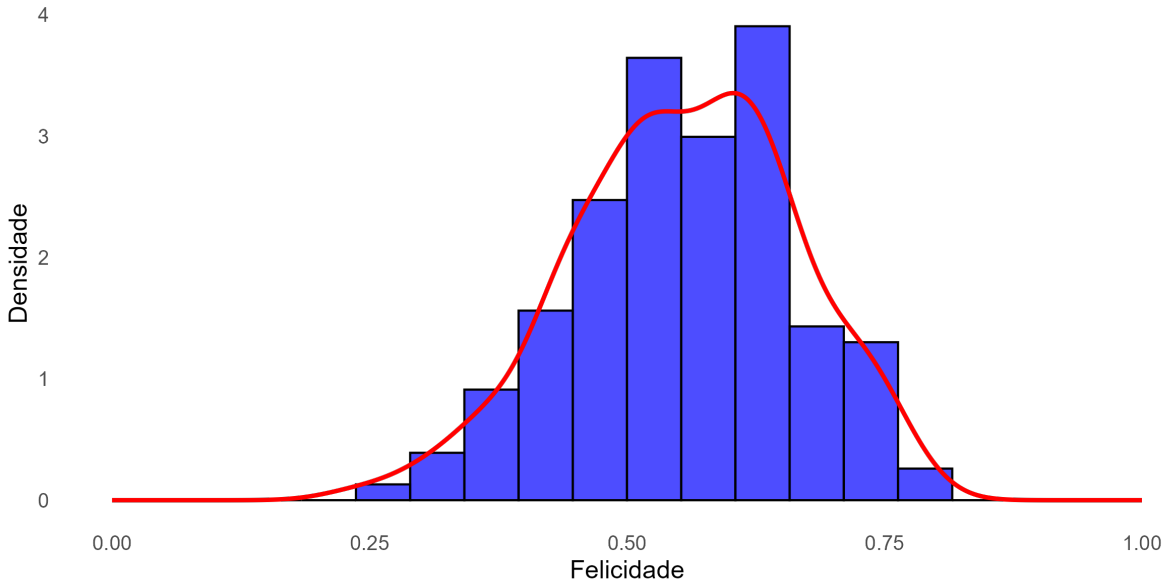
3 Resultados e Discussão

3.1 World Happines Report

O histograma e a curva de densidade apresentados na Figura 3 indicam que a variável felicidade apresenta uma distribuição concentrada em valores intermediários, sugerindo que a maioria das observações está situada próxima ao centro do intervalo com um leve deslocamento a direita. A densidade máxima ocorre nesse intervalo, demonstrando que os níveis de felicidade tendem a se agrupar em torno desse valor.

Além disso, a curva de densidade apresenta uma forma aproximadamente simétrica, sugerindo que os dados de felicidade não estão concentrados de maneira desproporcional em valores muito baixos ou muito altos.

Figura 3 – Histograma e densidade da variável resposta Felicidade

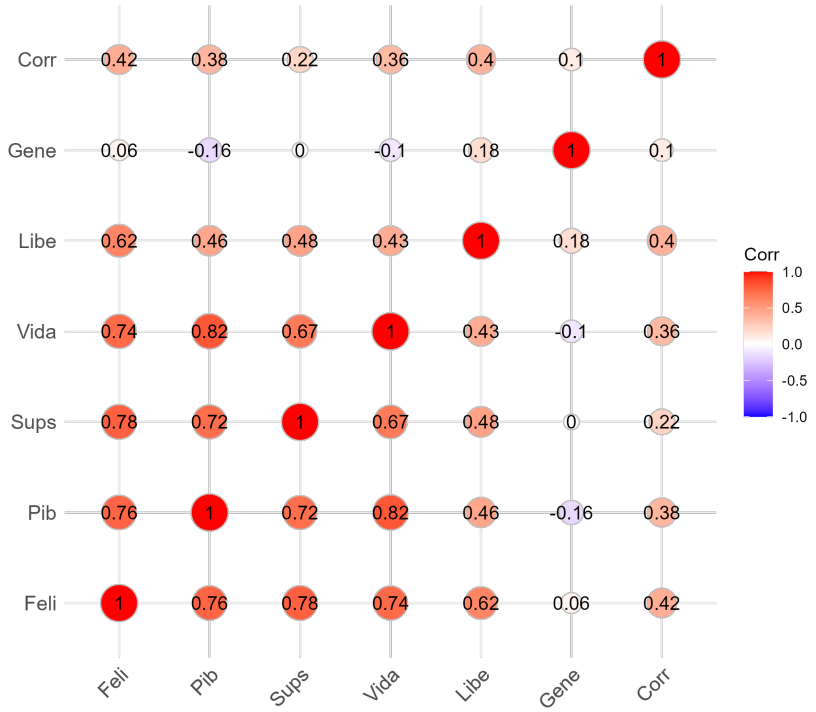


Para compreender os fatores que influenciam a felicidade, analisamos a correlação entre essa variável com as explicativas. A matriz de correlação apresentada na Figura 4 revelou relações significativas entre felicidade e essas variáveis, permitindo identificar padrões relevantes.

Os resultados indicam que Suporte Social (Sups) (0.78), Expectativa de Vida (Vida) (0.74) e PIB (0.76) apresentam as correlações mais fortes com felicidade, evidenciando que o suporte social, a expectativa de vida e o desenvolvimento econômico estão diretamente relacionados ao bem-estar da população.

A variável de confiança no governo (Corr) apresentou uma correlação positiva moderada com a felicidade, sugerindo que níveis reduzidos de corrupção institucional estão

Figura 4 – Correlação da variável Felicidade com as explicativas



associados a maior bem-estar populacional.

Outro destaque da matriz de correlação é a relação entre PIB e Expectativa de Vida (0.82), sugerindo que países com maior nível econômico tendem a oferecer condições que favorecem uma maior longevidade. Esse resultado pode refletir melhor acesso a serviços de saúde, educação e qualidade de vida, tornando o desenvolvimento econômico um fator crucial na promoção do bem-estar populacional. Já Liberdade para fazer Escolhas (Libe) apresentou correlação com outras variáveis importantes, indicando que ambientes onde os indivíduos possuem maior autonomia tendem a oferecer melhores condições para a qualidade de vida e fortalecimento das redes de suporte.

Foi calculado o Fator de Inflação de Variância (VIF, *Variance Inflation Factor*) para avaliar a presença de multicolinearidade entre os preditores. Os valores obtidos indicaram que todas as variáveis apresentam VIF abaixo do limite crítico de 5, indicando baixo risco de multicolinearidade severa. Embora PIB e Expectativa de Vida apresentem valores relativamente mais elevados (3.97 e 3.15, respectivamente), esses valores ainda estão dentro de uma faixa aceitável, permitindo sua inclusão no modelo sem comprometer a precisão das estimativas.

Foram testados os modelos propostos na seção 2.2, considerando a inclusão de todas as variáveis explicativas no parâmetro da média (μ). A comparação entre as distribuições foi realizada com base no AIC, permitindo identificar o modelo que melhor se ajusta aos dados.

Os resultados da análise (Tabela 3) indicaram que o modelo GB1 apresentou o menor valor de AIC, evidenciando que essa distribuição fornece o melhor ajuste estatístico

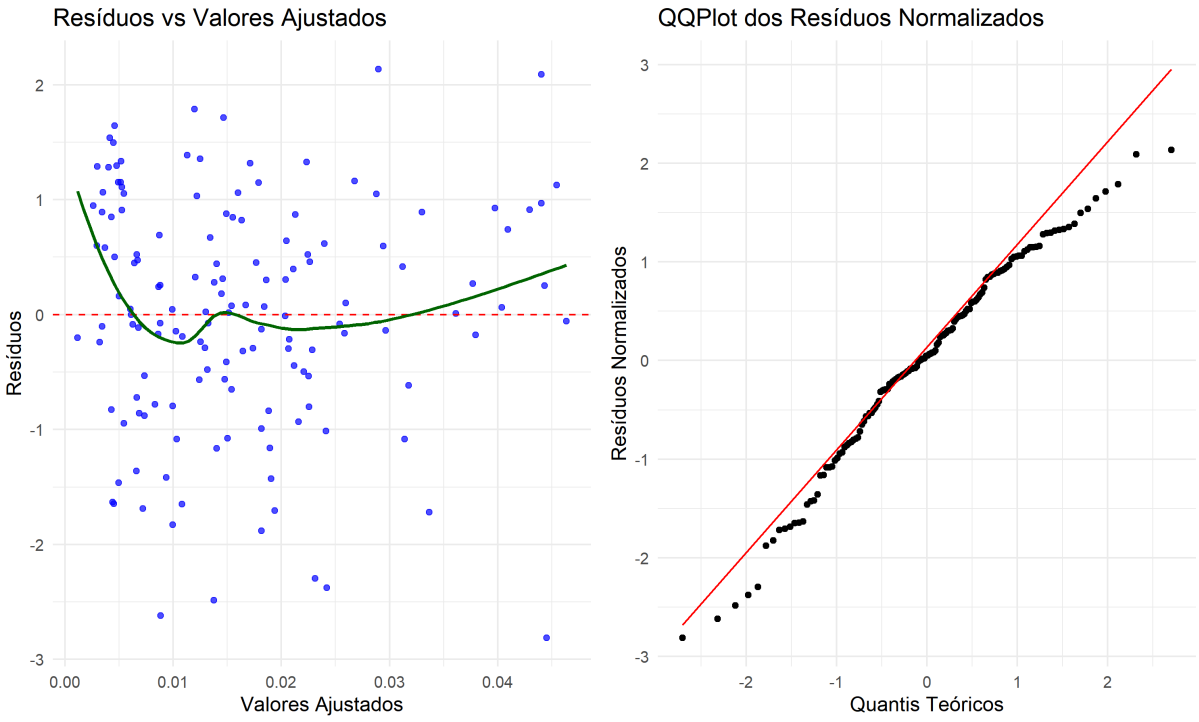
para a variável felicidade. Esse desempenho sugere que a distribuição Beta Generalizada Tipo 1 captura melhor a variabilidade e a estrutura dos dados, garantindo maior precisão nas estimativas dos coeficientes.

Tabela 3 – Comparação dos modelos ajustados com base no critério AIC

Modelo	**AIC**
Beta Generalizada Tipo 1	-457,66
Beta Generalizada Tipo 2	-444,67
Simplex	-439,23
Logito-Normal	-437,91
Beta	-435,92

Uma análise inicial dos resíduos exibidos na Figura 5 revela uma distribuição aproximadamente simétrica em torno de zero, sem indícios marcantes de *outliers*. No entanto, a presença de uma leve curvatura na linha suavizada indica um padrão, especialmente nos menores valores ajustados, sugerindo que o modelo pode não estar capturando totalmente a estrutura dos dados nessa região. Além disso, observa-se uma leve heterocedasticidade, com maior dispersão dos resíduos para valores ajustados baixos. A avaliação do QQPlot dos resíduos indica que, de modo geral, os pontos seguem a linha de referência nos quantis centrais, sugerindo uma aproximação razoável à normalidade nessa região. Pequenos desvios nas caudas evidenciam uma leve inadequação do modelo nos extremos da distribuição, mas sem indicar violações severas das suposições do modelo.

Figura 5 – Análise dos resíduos do modelo GB1



A seleção de variáveis do parâmetro de média (μ) foi realizada, permitindo avaliar a significância estatística de cada preditor no modelo GB1. Os resultados indicaram que a variável Generosidade (Gene) não apresentou significância estatística, conforme apresentado na Tabela 4, sugerindo que sua inclusão no modelo não contribui significativamente para explicar a variação da felicidade.

Tabela 4 – Análise de seleção de variáveis por meio da função drop1() no modelo GB1

Variável	**Df**	**AIC**	**LRT**	**Pr(Chi)**
Modelo completo	–	-457.66	–	–
Pib	1	-449.51	10.152	0.001442 **
Sups	1	-422.05	37.605	8.661e-10 ***
Vida	1	-451.08	8.584	0.003391 **
Libe	1	-437.65	22.006	2.718e-06 ***
Gene	1	-457.10	2.560	0.109605
Corr	1	-453.60	6.061	0.013822 *

Significância:
******* $p < 0.001$, ****** $p < 0.01$, ***** $p < 0.05$, . $p < 0.1$

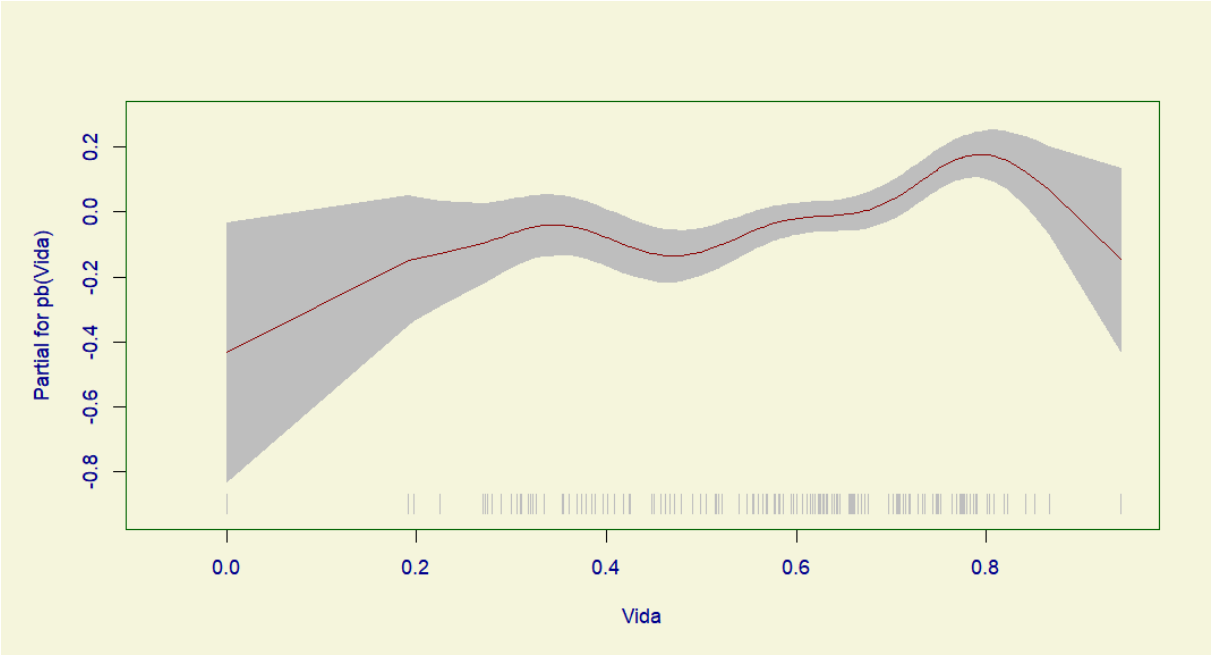
Além disso, a análise do AIC mostrou que a remoção de Generosidade (Gene) não alterou significativamente o valor do critério, indo de -457,66 com todas as covariáveis para -457,10 no modelo sem Generosidade, indicando que a exclusão dessa variável não compromete o ajuste do modelo. Com base nesses resultados, optamos por seguir com a nova configuração de variáveis explicativas, composta por Ausência de corrupção (Corr), Liberdade para fazer escolhas (Libe), Suporte Social (Sups), PIB e Expectativa de Vida (Vida).

Com o objetivo de testar o aumento da flexibilidade da modelagem e capturar possíveis padrões não lineares na relação entre as variáveis explicativas e a felicidade, foram adicionados suavizadores do tipo `pb()` (*Penalized B-splines*), que permitem modelar relações não lineares entre as covariáveis e a variável resposta de forma flexível. A adequação dessa abordagem foi testada por meio do AIC, permitindo comparar a qualidade do ajuste com e sem a suavização.

Os resultados indicaram que a inclusão do suavizador na variável Vida resultou em uma leve redução do AIC, passando de -457.10 para -459.03, o que representa uma melhoria marginal no ajuste do modelo. No entanto, ao analisar os gráficos gerados pelo *term plot* apresentado na Figura 6, observou-se que a suavização aplicada à variável Vida não apresentou uma estrutura clara de não linearidade, sugerindo que o efeito capturado pode estar mais relacionado a flutuações aleatórias do que a um padrão significativo nos dados. Diante desse cenário, optamos por remover o suavizador da variável, garantindo um modelo mais parcimonioso e reduzindo a possibilidade de *overfitting* a variações irrelevantes.

A seleção das variáveis que melhor explicam o parâmetro de dispersão (σ) foi realizada removendo preditores e comparando os modelos e o que apresentou melhor ajuste,

Figura 6 – Term plot da covariável Vida após a aplicação de suavizador do tipo pb()



com AIC reduzido para -463.19, inclui apenas PIB, Liberdade para fazer Escolhas (Libe) e Ausência de Corrupção (Corr), indicando que esses fatores não apenas afetam os níveis médios de felicidade, mas também sua variabilidade.

No modelo GAMLSS, o parâmetro τ é responsável por controlar a curtose da distribuição da variável resposta, ou seja, o peso das caudas e o grau de concentração dos valores em torno da média. Após seleção das variáveis deste parâmetro, o modelo que apresentou melhor ajuste incluiu apenas PIB e e Suporte Social (Sups), sugerindo que esses fatores desempenham um papel relevante na modelagem da curtose.

O parâmetro ν está relacionado à assimetria da distribuição. Na etapa de seleção de variáveis para esse parâmetro, nenhuma das covariáveis testadas apresentou significância estatística, indicando que não há preditores que contribuam de forma relevante para explicar possíveis assimetrias na variável felicidade. Diante desse resultado, optou-se por manter o modelo final sem incluir termos específicos para ν, garantindo uma especificação mais parcimoniosa e focada nos parâmetros que efetivamente influenciam a distribuição da resposta.

Após a seleção inicial das variáveis com o uso da função `drop1()`, que prioriza o ajuste global do modelo com base no AIC, procedeu-se à avaliação dos coeficientes individuais por meio do `summary()`. Essa etapa foi fundamental para identificar preditores que, embora contribuíssem para a adequação geral do modelo, não apresentavam significância estatística robusta isoladamente. Dessa forma, a combinação das abordagens permitiu não apenas otimizar o ajuste do modelo segundo o AIC, mas também aprimorar a interpretação dos efeitos individuais dos preditores.

O resultado do `summary()` pode ser observado abaixo:

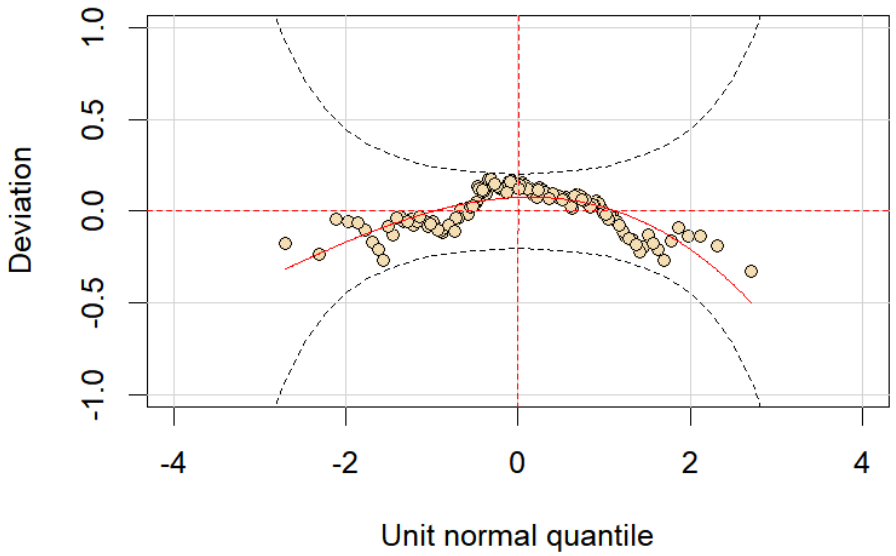
```
******************************************************************
Family:  c("GB1", "Generalized beta type 1")

Call:  gamlss(formula = Feli ~ Pib + Vida + Sups + Libe +
    Corr, sigma.formula = ~Pib + Libe + Corr, tau.formula = ~Pib +
    Sups, family = GB1, data = dados_feli, method = RS(10000))

Fitting method: RS(10000)

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.9462     0.9424  -6.309 3.92e-09 ***
Pib           1.3760     0.3810   3.611 0.000432 ***
Vida          0.6054     0.2365   2.560 0.011604 *
Sups          1.4174     0.4508   3.144 0.002056 **
Libe          0.7958     0.2160   3.684 0.000333 ***
Corr          0.2062     0.1541   1.339 0.182991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  logit
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.92848    0.84015  -3.486 0.000667 ***
Pib          0.01753    0.32657   0.054 0.957274
Libe         0.08041    0.46884   0.172 0.864092
Corr         0.90190    0.54173   1.665 0.098314 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  log
Nu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.218      1.315   1.686   0.0941 .
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Tau link function:  log
Tau Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6543     0.3180   5.202 7.35e-07 ***
Pib          -0.8635     0.4445  -1.943   0.0542 .
Sups         -0.5967     0.4046  -1.475   0.1427
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
No. of observations in the fit:  146
Degrees of Freedom for the fit:  14
      Residual Deg. of Freedom:  132
                      at cycle:  3556

Global Deviance:     -491.2938
            AIC:     -463.2938
            SBC:     -421.5233
******************************************************************
```

Ao avaliar o AIC do modelo resultante dessa seleção adicional, constatou-se uma melhoria significativa. Em virtude desse aprimoramento, optou-se por remover variáveis que não apresentaram relevância estatística consistente: a variável de Ausência de Corrupção (Corr) foi excluída da modelagem do parâmetro de media (μ), as variáveis PIB e Liberdade para fazer Escolhas (Libe) foram retiradas da formulação do parâmetro de dispersão (σ) e a variável Suporte Social (Sups) deixando de compor o parâmetro de curtose(τ). Essa estratégia resultou em um modelo final mais parcimonioso e interpretável, com melhor desempenho global conforme evidenciado pela redução do AIC que foi de -463.30 para -467.55.

O *worm plot* apresentado na Figura 7 mostra que a maior parte dos resíduos permanece dentro das bandas de confiança, o que sugere que o modelo apresenta um ajuste global satisfatório. No entanto, a curva suavizada (linha vermelha) apresenta uma leve oscilação, com desvios negativos nas caudas e positivos no centro, indicando um pequeno excesso de curtose nos resíduos normalizados. Ainda assim, como os desvios não ultrapassam os limites críticos, não há indícios de inadequação severa.

Figura 7 – Worm plot dos resíduos normalizados do modelo com distribuição GB1



A seguir, apresenta-se o modelo final ajustado, considerando as funções de ligação para cada parâmetro da distribuição GB1:

- **Parâmetro de Média** (μ)

A seguir, apresenta-se a equação que relaciona o logito da média da variável resposta com as covariáveis do modelo:

$$\text{logit}(\mu) = -6{,}369 + 1{,}759 \times \text{Pib} + 0{,}645 \times \text{Vida} + 0{,}637 \times \text{Sups} + 0{,}927 \times \text{Libe}$$

Essa formulação indica que o logito da média da variável resposta aumenta linearmente com as covariáveis PIB, Expectativa de Vida (Vida), Suporte Social (Sups) e Liberdade de Escolha (Libe). Como a função logit é a transformação logarítmica da razão entre μ e $1-\mu$, os coeficientes podem ser interpretados em termos das odds da média.

Em particular, para cada aumento unitário no PIB, a log-odds de μ aumenta em 1,7590, o que equivale a multiplicar as odds de μ por aproximadamente $e^{1{,}759} \approx 5,80$. Isso significa que, mantendo as demais variáveis constantes, um aumento no PIB está associado a um aumento expressivo no valor esperado da variável resposta. De maneira semelhante, as variáveis Esxpectativa de Vida (Vida), Suporte Social (Sups) e Liberdade de Escolha (Libe) também apresentam efeitos positivos, indicando que maiores valores nessas covariáveis estão relacionados a aumentos na média da resposta.

- **Parâmetro de Dispersão** (σ)

A seguir, apresenta-se a equação que descreve a relação entre o logito da dispersão da variável resposta e as covariáveis incluídas no modelo:

$$\text{logit}(\sigma) = -3{,}003 + 1{,}151 \times \text{Corr}$$

Para cada unidade de aumento em Ausência de Corrupção (Corr), o logito de σ aumenta em 1,151 unidades, ou seja, a odds de σ aumenta em $e^{1,1508} \approx 3,16$ vezes.

- **Parâmetro de Curtose (τ)**

A seguir, apresenta-se a equação que descreve a relação entre o log da curtose da variável resposta e as covariáveis incluídas no modelo:

$$\ln(\tau) = 1{,}640 - 1{,}019 \times \text{Pib}$$

Como τ está relacionado à curtose (peso das caudas da distribuição da variável resposta), essa relação sugere que países com maior PIB tendem a apresentar distribuições com caudas mais leves, refletindo uma menor probabilidade de ocorrência de valores extremos na variável dependente. Em contrapartida, países com PIB mais baixo estariam mais associados a distribuições com maior presença de valores extremos.

- **Parâmetro de Assimetria (ν)**

A seguir, apresenta-se a equação que descreve a relação entre o log da assimetria da variável resposta e as covariáveis incluídas no modelo:

$$\ln(\nu) = 2{,}326$$

O uso de apenas o intercepto sugere que o modelo não identificou variáveis explicativas que influenciem de forma estatisticamente significativa a assimetria da distribuição da variável resposta. Dessa forma, o valor de ν é fixo para todas as observações, assumindo um efeito constante sobre a assimetria ao longo de toda a amostra.

O resultado do summary() do modelo final ajustado segue:

```
******************************************************************
Family:  c("GB1", "Generalized beta type 1")

Call:  gamlss(formula = Feli ~ Pib + Vida + Sups + Libe, sigma.formula = ~Corr,
    tau.formula = ~Pib, family = GB1, data = dados_feli,      method = RS(10000))

Fitting method: RS(10000)

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3696     1.1456  -5.560 1.38e-07 ***
```

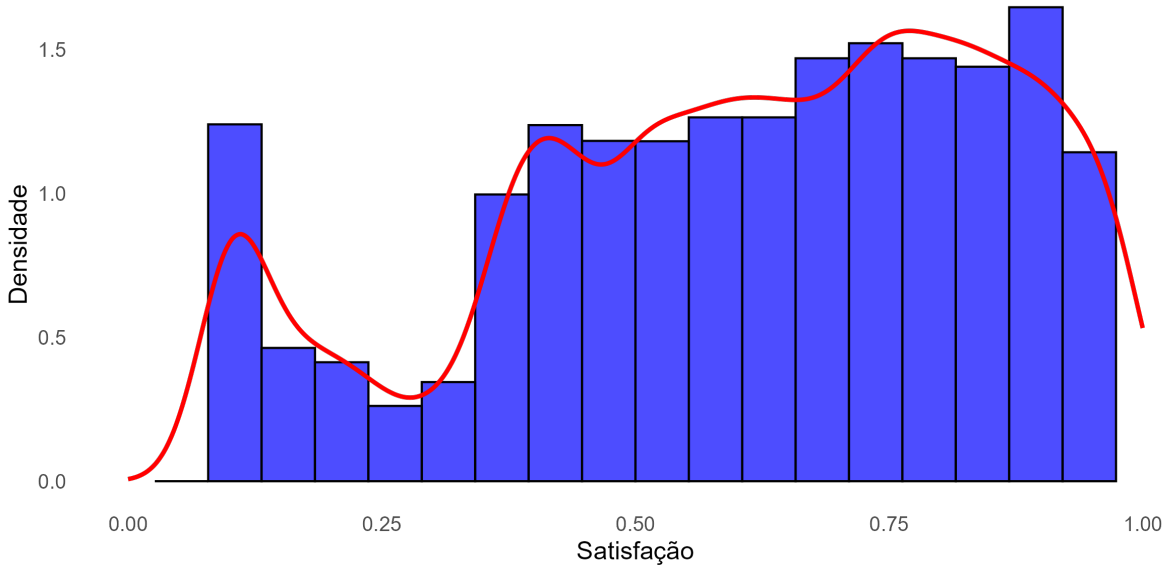
```
Pib           1.7590     0.2816   6.246 5.03e-09 ***
Vida          0.6450     0.2282   2.827 0.005416 **
Sups          0.6368     0.1869   3.408 0.000862 ***
Libe          0.9275     0.2032   4.565 1.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  logit
Sigma Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.0033     0.4001  -7.507 7.16e-12 ***
Corr          1.1508     0.5464   2.106    0.037 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  log
Nu Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3265     0.8468   2.747  0.00682 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Tau link function:  log
Tau Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6401     0.2700   6.074 1.18e-08 ***
Pib          -1.0190     0.3572  -2.853  0.00501 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
No. of observations in the fit:  146
Degrees of Freedom for the fit:  10
      Residual Deg. of Freedom:  136
                      at cycle:  3133
```

```
Global Deviance:     -487.5552
            AIC:     -467.5552
            SBC:     -437.7191
************************************************************
```

3.2 Employee Satisfaction Survey Data

O histograma e a curva de densidade apresentados na Figura 8 indicam que a variável satisfação apresenta assimetria à esquerda, indicando que a maior concentração de empregados possui níveis mais altos de satisfação. A curva de densidade (em vermelho) reforça essa observação ao evidenciar dois picos principais: um menor, em torno de 0,1 a 0,15, e outro mais pronunciado, entre 0,7 e 0,9. Essa configuração sugere uma possível distribuição bimodal. Além disso, nota-se uma baixa frequência de respostas intermediárias, o que pode indicar uma polarização nas opiniões dos empregados quanto à sua satisfação no ambiente de trabalho.

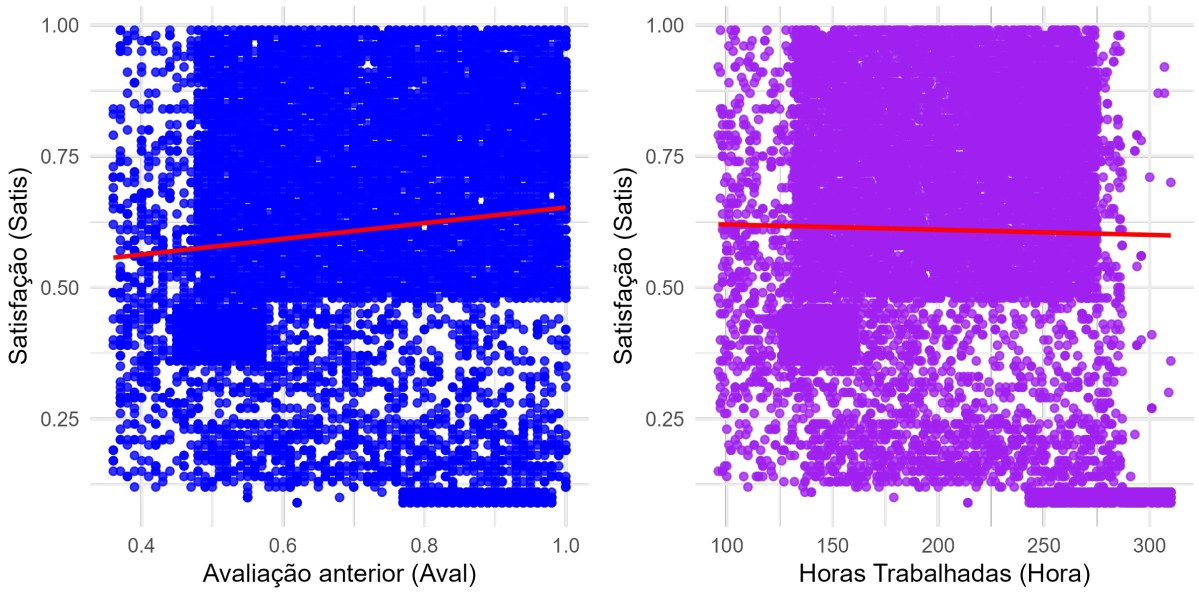
Figura 8 – Histograma e densidade da variável resposta Satisfação



Os gráficos da Figura 9 mostram a relação da resposta (Satisfação do empregados) com as covariáveis numéricas. Para a variável Avaliação Anterior (Aval), os pontos azuis indicam que há uma tendência de que níveis mais altos estejam associados a maiores valores de Satisfação (Sati). A linha de tendência (em vermelho) reforça essa associação positiva, ainda que de forma moderada. Esse padrão sugere que os funcionários que tiveram avaliações anteriores mais favoráveis tendem a relatar uma maior satisfação. Essa relação pode refletir uma experiência prévia positiva, que se traduz em uma melhor percepção do ambiente ou das condições de trabalho.

Para a variável Horas trabalhadas (Hora), a linha de tendência em vermelho indica uma relação negativa, à medida que o número de horas trabalhadas aumenta, a satisfação tende a diminuir ligeiramente. Esse padrão pode sugerir que jornadas mais longas ou cargas altas de trabalho estão associadas a um declínio na percepção de satisfação dos funcionários, possivelmente devido a fatores como fadiga, estresse ou desequilíbrio entre vida pessoal e profissional.

Figura 9 – Relação das variáveis explicativas numéricas com a resposta Satisfação

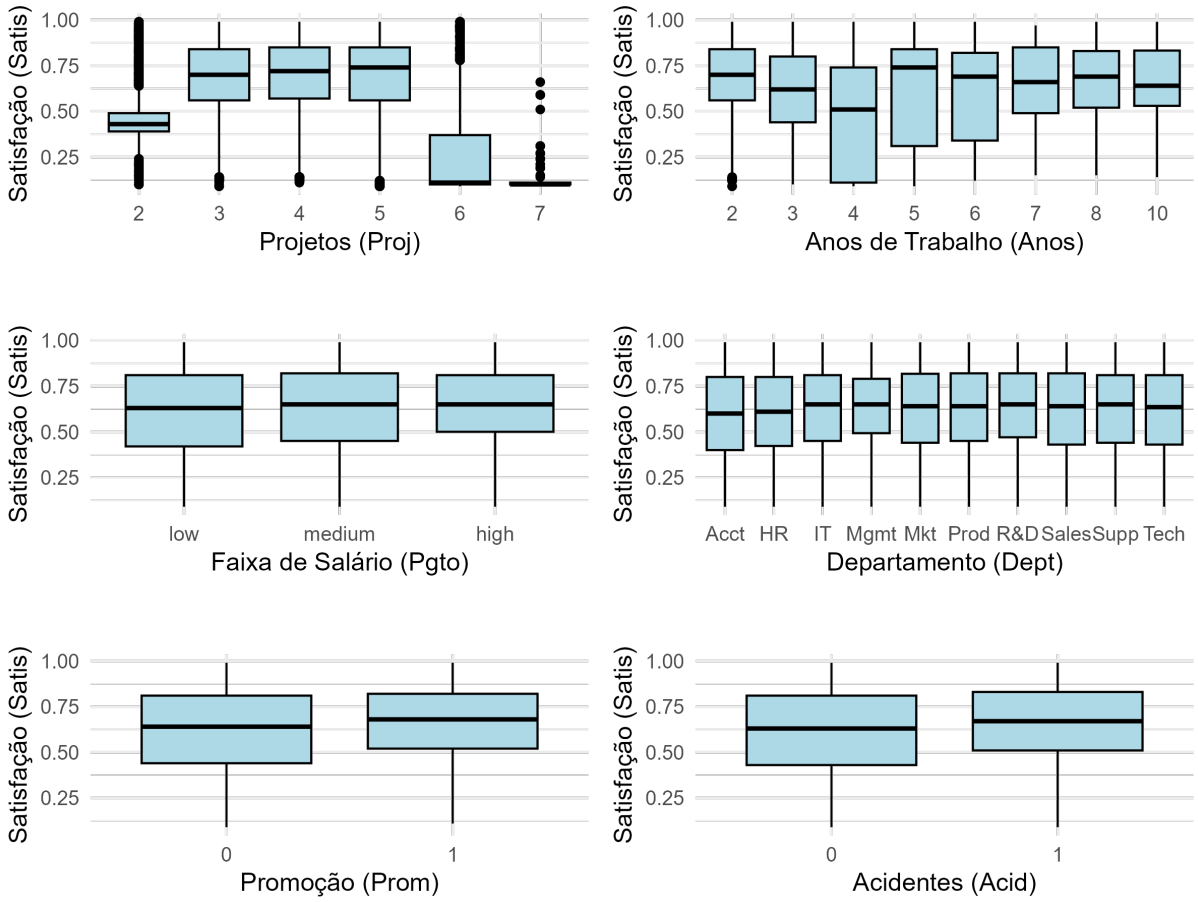


Os gráficos da Figura 10 mostram a relação da resposta com as covariáveis categóricas e os boxplots sugerem que há variações na satisfação de acordo com a quantidade de projetos em que os empregados estão envolvidos. Pode-se notar diferenças na mediana e na dispersão entre os níveis de projetos. Em alguns grupos, a variação interquartil pode ser maior, indicando maior heterogeneidade na satisfação. Essa análise sugere que o número de projetos pode influenciar significativamente a percepção de satisfação, seja por sobrecarga ou por engajamento variável.

Observa-se que, de maneira geral, a variação da satisfação ao longo dos anos de serviço é relativamente consistente, com medianas próximas entre as diferentes categorias. Contudo, pequenos desvios ou variações nos intervalos interquartis podem indicar que, para determinados períodos de tempo, os sentimentos dos empregados divergem um pouco se os empregados mais antigos têm uma distribuição levemente mais concentrada ou dispersa, o que pode refletir adaptação ou desgaste ao longo do tempo.

As categorias de faixa de pagamento, departamento, promoção e ocorrência de acidentes exibem distribuições muito semelhantes, com medianas e dispersões praticamente idênticas, sugerindo que esses fatores isoladamente não influenciam de forma significativa a percepção de satisfação dos empregados.

Figura 10 – Relação das variáveis explicativas categóricas com a resposta Satisfação



Após avaliar a multicolinearidade entre as variáveis explicativas incluídas no modelo por meio do cálculo do VIF, constatamos que todos os valores se encontraram abaixo do limiar de 5. Esse resultado indica uma baixa multicolinearidade, sugerindo que as variáveis independentes não apresentam dependência linear relevante e que as estimativas dos coeficientes da regressão podem ser obtidas de forma consistente.

Foram testados os modelos propostos na seção 2.2, considerando a inclusão de todas as variáveis explicativas no parâmetro de média (μ). A comparação entre as distribuições foi realizada com base no AIC, permitindo identificar o modelo que melhor se ajusta aos dados.

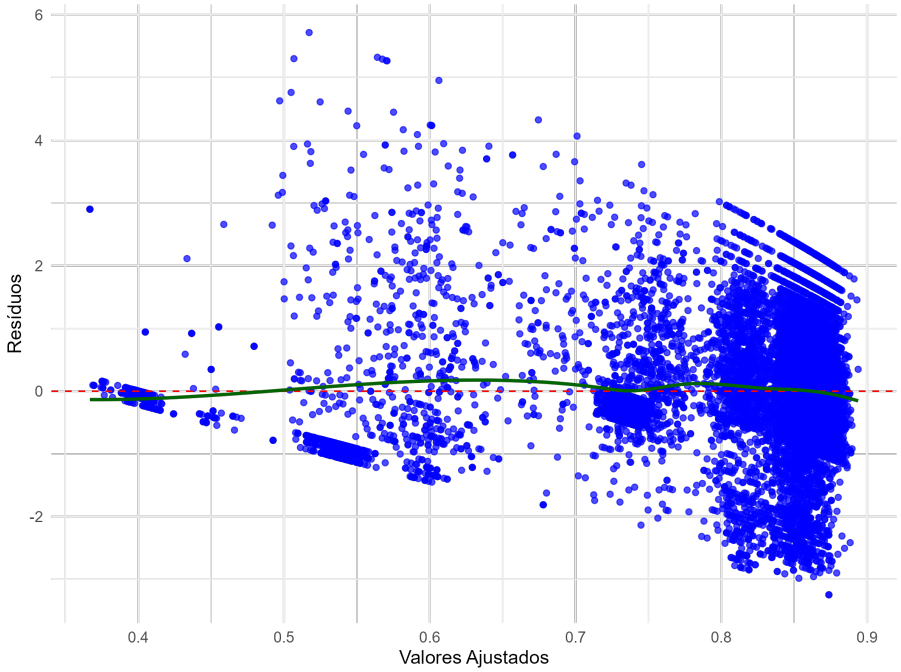
Os resultados da análise (Tabela 5) indicaram que o modelo GB1 apresentou o menor valor de AIC, evidenciando que essa especificação fornece o melhor ajuste estatístico para a variável satisfação.

Tabela 5 – Comparação dos modelos ajustados com base no critério AIC

Modelo	**AIC**
Beta Generalizada Tipo 1	-8783,318
Logito-Normal	-8486.746
Beta	-8662,254
Beta Generalizada Tipo 2	-5930,210
Simplex	-5379,314

A análise gráfica dos resíduos versus valores ajustados observado na Figura 11 indicou a presença de heterocedasticidade, evidenciada pela dispersão desigual dos resíduos ao longo dos valores ajustados. Além disso, verificou-se a existência de resíduos padronizados fora da faixa entre -3 e 3, e uma concentração excessiva de pontos em algumas faixas de valores ajustados.

Figura 11 – Gráfico de resíduos versus Valores Ajustados do modelo GB1



A seleção de variáveis do parâmetro μ foi realizada por meio da função `drop1()`, permitindo avaliar a significância estatística de cada preditor no modelo GB1. Os resultados indicaram que as variáveis significativas são apenas Avaliação Anterior (Aval), Quantidade de Projetos (Proj), Anos nas Empresa (Anos) e Departamento (Dept). A análise do AIC mostrou melhora significativa na nova configuração das variáveis, indicando que a exclusão não compromete o ajuste do modelo.

Com o objetivo de aumentar a flexibilidade da modelagem e capturar possíveis relações não lineares entre os preditores e a variável resposta, foi testada a inclusão de um suavizador do tipo `pb()`, que penaliza a oscilação, na variável Avaliação Anterior (Aval). Essa decisão se justifica pelo fato de Aval ser uma variável numérica contínua, o que permite a aplicação de suavizadores para estimar efeitos de forma mais flexível.

Por outro lado, as demais variáveis preditoras foram mantidas em sua forma original, sem aplicação de suavizadores, pois tratam-se de variáveis categóricas. Nesse caso, a inclusão de funções de suavização não é apropriada, uma vez que não há uma ordem ou estrutura contínua entre os níveis dessas variáveis que justifique a suavização.

A introdução do suavizador `pb()` na variável Aval resultou em uma melhora no ajuste do modelo, refletida na redução do AIC, que passou para -8862.477. Esse resultado reforça a hipótese de que a relação entre Aval e a variável resposta apresenta características não lineares, sendo mais adequadamente modelada por meio de uma função de suavização.

A modelagem do parâmetro σ revelou que a heterogeneidade nos dados está associada a diversas covariáveis. O processo de seleção de variáveis, indicou que as variáveis Avaliação Anterior (Aval), Número de Projetos (Proj), Horas Trabalhadas (Hora), Anos de empresa (Anos), Acidente de Trabalho (Acid), Departamento (Dept) e Faixa Salarial (Pgto) contribuem de forma significativa para explicar a variabilidade presente na resposta com um ganho no AIC (-11032,867). Isso sugere que essas covariáveis influenciam diretamente a dispersão dos níveis de satisfação observados, refletindo diferenças na consistência das respostas entre os diferentes perfis analisados.

Para o parâmetro τ, observou-se que as variáveis Número de Projetos (Proj) e Acidente de Trabalho (Acid) foram estatisticamente significativas. Isso indica que tanto a quantidade de projetos atribuídos ao indivíduo quanto a ocorrência de acidentes de trabalho contribuem para explicar a variação na curtose da distribuição da variável resposta. Em outras palavras, esses fatores influenciam a concentração e o peso das caudas da distribuição, sugerindo que a forma da distribuição da satisfação é sensível a essas características.

No parâmetro ν, responsável por controlar a assimetria da distribuição, apenas a variável Quantidade de Projetos (Proj) teve significância estatística, indicando que apenas a quantidade de projetos interfere na assimetria das respostas.

Após a seleção das variáveis, foi realizada a avaliação dos coeficientes individuais por meio do `summary()` para identificar preditores que não apresentavam significância estatística isolados. Assim, foi possível combinar as abordagens para otimizar o ajuste e interpretação dos efeitos individuais.

O resultado do `summary()` pode ser observado abaixo:

```
******************************************************************
Family:  c("GB1", "Generalized beta type 1")

Call:  gamlss(formula = Sati ~ pb(Aval) + Proj + Anos + Dept,
sigma.formula = ~Aval + Proj + Hora + Anos + Acid + Dept +
Pgto, nu.formula = ~Proj, tau.formula = ~Proj + Acid,
family = GB1, data = dados, method = RS(10000))
```

```
Fitting method: RS(10000)

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.60608    0.06235   9.720  < 2e-16 ***
pb(Aval)         0.59894    0.05911  10.133  < 2e-16 ***
Proj3           -1.57524    0.03401 -46.319  < 2e-16 ***
Proj4           -2.00844    0.03373 -59.540  < 2e-16 ***
Proj5           -1.49208    0.03516 -42.436  < 2e-16 ***
Proj6            6.44491    0.04402 146.413  < 2e-16 ***
Proj7            4.70214    0.06760  69.557  < 2e-16 ***
Anos3           -0.15220    0.02499  -6.091 1.15e-09 ***
Anos4           -0.63018    0.03170 -19.882  < 2e-16 ***
Anos5           -0.31549    0.03597  -8.771  < 2e-16 ***
Anos6           -0.32294    0.04774  -6.765 1.38e-11 ***
Anos7           -0.14502    0.08749  -1.657  0.09744 .
Anos8           -0.14282    0.09116  -1.567  0.11719
Anos10          -0.13343    0.08421  -1.584  0.11311
Depthr           0.03998    0.05748   0.696  0.48669
DeptIT           0.11574    0.05099   2.270  0.02324 *
Deptmanagement   0.08348    0.06027   1.385  0.16603
Deptmarketing    0.12166    0.05554   2.190  0.02851 *
Deptproduct_mng  0.15578    0.05500   2.832  0.00463 **
DeptRandD        0.08391    0.05684   1.476  0.13990
Deptsales        0.10824    0.04390   2.466  0.01369 *
Deptsupport      0.11035    0.04661   2.367  0.01793 *
Depttechnical    0.12777    0.04557   2.804  0.00505 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  logit
Sigma Coefficients:
                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)     2.3107135  0.0574714   40.206  < 2e-16 ***
Aval           -0.2487947  0.0471884   -5.272 1.37e-07 ***
Proj3          -1.2372971  0.0263281  -46.995  < 2e-16 ***
```

```
Proj4           -1.2584201  0.0262298  -47.977  < 2e-16 ***
Proj5           -1.8145107  0.0291285  -62.293  < 2e-16 ***
Proj6           -4.2584600  0.0310848 -136.995  < 2e-16 ***
Proj7           -3.9315119  0.0506630  -77.601  < 2e-16 ***
Hora            -0.0010584  0.0001614   -6.557 5.67e-11 ***
Anos3           -0.1642889  0.0202688   -8.106 5.66e-16 ***
Anos4            0.0019394  0.0246422    0.079  0.93727
Anos5            0.1571512  0.0282556    5.562 2.72e-08 ***
Anos6            0.3342864  0.0359886    9.289  < 2e-16 ***
Anos7            0.1017429  0.0673348    1.511  0.13081
Anos8            0.0083394  0.0719539    0.116  0.90773
Anos10           0.1714284  0.0636039    2.695  0.00704 **
Acid1            0.1089788  0.0206641    5.274 1.35e-07 ***
Depthr          -0.0731924  0.0446583   -1.639  0.10125
DeptIT          -0.0635251  0.0397710   -1.597  0.11023
Deptmanagement  -0.0659015  0.0475171   -1.387  0.16549
Deptmarketing   -0.0272651  0.0432365   -0.631  0.52831
Deptproduct_mng  0.0314582  0.0424717    0.741  0.45890
DeptRandD        0.0687604  0.0440664    1.560  0.11869
Deptsales        0.0102674  0.0339615    0.302  0.76241
Deptsupport     -0.0239424  0.0361856   -0.662  0.50820
Depttechnical    0.0120979  0.0352032    0.344  0.73111
Pgtolow         -0.1314289  0.0279776   -4.698 2.66e-06 ***
Pgtomedium      -0.0830808  0.0280485   -2.962  0.00306 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-----------------------------------------------------------------
Nu link function:  log
Nu Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -3.605e+01  1.426e-01 -252.820   <2e-16 ***
Proj3        3.489e+01  1.467e-01  237.782   <2e-16 ***
Proj4        3.536e+01  1.469e-01  240.654   <2e-16 ***
Proj5        3.712e+01  1.450e-01  256.014   <2e-16 ***
Proj6       -7.735e-03  1.632e-01   -0.047    0.962
Proj7       -7.232e-04  1.671e-01   -0.004    0.997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
-------------------------------------------------------------------
Tau link function:  log
Tau Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.652278   0.004065  898.51  < 2e-16 ***
Proj3       -1.799827   0.012948 -139.00  < 2e-16 ***
Proj4       -1.627712   0.013161 -123.67  < 2e-16 ***
Proj5       -3.158415   0.019026 -166.00  < 2e-16 ***
Proj6       -1.114138   0.006223 -179.03  < 2e-16 ***
Proj7       -1.069144   0.005308 -201.41  < 2e-16 ***
Acid1        0.024412   0.007138    3.42 0.000628 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-------------------------------------------------------------------
NOTE: Additive smoothing terms exist in the formulas:
 i) Std. Error for smoothers are for the linear effect only.
ii) Std. Error for the linear terms may not be reliable.
-------------------------------------------------------------------
No. of observations in the fit:  14888
Degrees of Freedom for the fit:  71.44986
      Residual Deg. of Freedom:  14816.55
                      at cycle:  2735

Global Deviance:     -14705.45
            AIC:     -14562.55
            SBC:     -14018.94
*******************************************************************
```

Observou-se que todos os níveis da variável Departamento (Dept) apresentaram coeficientes não significativos para o parâmetro σ, indicando que essa variável não exerce efeito estatisticamente relevante sobre a dispersão da resposta. Como o AIC desta nova modelagem não teve alteração significativa, optou-se por prosseguir sem a variável.

Dessa forma, o modelo final ajustado, considerando as funções de ligação para cada parâmetro da distribuição GB1 segue:

- **Parâmetro de Média** (μ)

A seguir, apresenta-se a equação que relaciona o logito da média da variável resposta com as covariáveis do modelo:

$$\begin{aligned}\text{logit}(\mu) = {} & 0{,}603 + f(\text{Aval}) - 1{,}564 \times \text{Proj3} - 2{,}018 \times \text{Proj4} - 1{,}496 \times \text{Proj5} \\ & + 6{,}384 \times \text{Proj6} + 4{,}586 \times \text{Proj7} - 0{,}152 \times \text{Anos3} - 0{,}633 \times \text{Anos4} \\ & - 0{,}313 \times \text{Anos5} - 0{,}323 \times \text{Anos6} - 0{,}143 \times \text{Anos7} - 0{,}139 \times \text{Anos8} \\ & - 0{,}135 \times \text{Anos10} + 0{,}047 \times \text{Depthr} + 0{,}121 \times \text{DeptIT} \\ & + 0{,}090 \times \text{Deptmanagement} + 0{,}125 \times \text{Deptmarketing} \\ & + 0{,}156 \times \text{Deptproduct\_mng} + 0{,}075 \times \text{DeptRandD} + 0{,}108 \times \text{Deptsales} \\ & + 0{,}112 \times \text{Deptsupport} + 0{,}128 \times \text{Depttechnical}\end{aligned}$$

A variável Avaliação (Aval) foi modelada com um efeito suavizado, permitindo capturar uma relação não-linear entre a avaliação de desempenho do colaborador e a média da variável resposta.

No caso da variável Proj, que representa a quantidade de projetos nos quais o colaborador está envolvido, observou-se que participações em um número reduzido de projetos estão associadas a uma diminuição do logito da média da variável resposta, enquanto quantidades mais elevadas de projetos estão relacionadas a um aumento expressivo no logito da média.

Quanto ao tempo de empresa (Anos), os resultados indicaram que colaboradores com 4 a 6 anos de empresa apresentaram uma redução mais acentuada no logito da média da variável resposta, quando comparados tanto aos colaboradores com 3 anos quanto àqueles com mais de 7 anos de empresa.

Os diferentes departamentos também apresentaram influência significativa. Em comparação ao departamento de referência, todos os demais setores analisados mostraram coeficientes positivos, sugerindo que, de forma geral, colaboradores de áreas como Product Management Support, Technical e Management possuem, em média, valores mais elevados da variável resposta após a transformação logística.

- **Parâmetro de Dispersão (σ)**

A seguir, apresenta-se a equação que descreve a relação entre o logito da dispersão da variável resposta e as covariáveis incluídas no modelo:

$$\begin{aligned}\text{logit}(\sigma) = {} & 2{,}295 - 0{,}249 \times \text{Aval} - 1{,}247 \times \text{Proj3} - 1{,}258 \times \text{Proj4} - 1{,}814 \times \text{Proj5} \\ & - 4{,}223 \times \text{Proj6} - 3{,}862 \times \text{Proj7} - 0{,}001 \times \text{Hora} - 0{,}164 \times \text{Anos3} \\ & + 0{,}004 \times \text{Anos4} + 0{,}155 \times \text{Anos5} + 0{,}337 \times \text{Anos6} + 0{,}098 \times \text{Anos7} \\ & + 0{,}009 \times \text{Anos8} + 0{,}158 \times \text{Anos10} + 0{,}112 \times \text{Acid1} \\ & - 0{,}121 \times \text{Pgtolow} - 0{,}072 \times \text{Pgtomedium}\end{aligned}$$

O efeito da variável Avaliação Anterior (Aval), com coeficiente negativo, indica que colaboradores com maior avaliação de desempenho tendem a apresentar menor dispersão

na variável resposta. Em relação ao número de projetos (Proj), os coeficientes foram todos negativos e esses resultados indicam que, à medida que o número de projetos aumenta, há uma redução expressiva na dispersão da variável resposta. Isso sugere que os colaboradores com maior carga de projetos tendem a apresentar respostas mais concentradas..

A variável Horas Trabalhadas (Hora) apresentou um coeficiente muito próximo de zero, indicando um efeito praticamente nulo sobre a dispersão. Quanto à variável tempo de empresa (Anos), os coeficientes para o parâmetro de dispersão apresentaram sinais variados. Esse padrão não segue uma tendência clara com o tempo de empresa, o que dificulta uma interpretação mais objetiva desse efeito.

Além disso, a variável Acidente de Trabalho (Acid), apresentou um coeficiente positivo, indicando aumento da dispersão entre os colaboradores que já tiveram acidente de trabalho.

Por outro lado, os níveis de salário baixo e médio mostraram diminuição na dispersão quando comparados ao alto.

- **Parâmetro de Curtose (τ)**

A seguir, apresenta-se a equação que descreve a relação entre o log da curtose da variável resposta e as covariáveis incluídas no modelo:

$$\begin{aligned}\log(\tau) = 3{,}652 - 1{,}812 \times \text{Proj3} - 1{,}618 \times \text{Proj4} - 3{,}148 \times \text{Proj5} \\ - 1{,}112 \times \text{Proj6} - 1{,}066 \times \text{Proj7} + 0{,}023 \times \text{Acid1}\end{aligned}$$

Os resultados indicaram que a quantidade de projetos tem forte influência na curtose da variável resposta, porém não foi observado um padrão claro, onde a curtose diminuísse progressivamente com o aumento no número de projetos, o que foge de uma expectativa de comportamento linear ou gradual com base na quantidade de projetos.

- **Parâmetro de Assimetria (ν)**

A seguir, apresenta-se a equação que descreve a relação entre o log da assimetria da variável resposta e as covariáveis incluídas no modelo:

$$\begin{aligned}\log(\nu) = -36{,}05 + 34{,}90 \times \text{Proj3} + 35{,}36 \times \text{Proj4} + 37{,}12 \times \text{Proj5} \\ - 0{,}008 \times \text{Proj6} - 0{,}001 \times \text{Proj7}\end{aligned}$$

De maneira geral, os resultados mostram que a quantidade de projetos tem um impacto relevante na assimetria da variável resposta, porém com um padrão de efeito bastante concentrado nas faixas intermediárias de número de projetos, e praticamente nulo nas faixas mais altas.

O resultado do summary() do modelo final ajustado segue:

```
******************************************************************
Family:  c("GB1", "Generalized beta type 1")

Call:  gamlss(formula = Sati ~ pb(Aval) + Proj + Anos + Dept,
sigma.formula = ~ Aval + Proj + Hora + Anos + Acid + Pgto,
nu.formula = ~Proj, tau.formula = ~Proj + Acid,
family = GB1, data = dados, method = RS(10000))


Fitting method: RS(10000)

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.60322    0.06233   9.678  < 2e-16 ***
pb(Aval)         0.59854    0.05910  10.127  < 2e-16 ***
Proj3           -1.56386    0.03401 -45.988  < 2e-16 ***
Proj4           -2.01804    0.03374 -59.807  < 2e-16 ***
Proj5           -1.49642    0.03519 -42.528  < 2e-16 ***
Proj6            6.38364    0.04403 144.975  < 2e-16 ***
Proj7            4.58578    0.06775  67.684  < 2e-16 ***
Anos3           -0.15213    0.02499  -6.088 1.17e-09 ***
Anos4           -0.63270    0.03170 -19.957  < 2e-16 ***
Anos5           -0.31324    0.03597  -8.708  < 2e-16 ***
Anos6           -0.32327    0.04775  -6.770 1.34e-11 ***
Anos7           -0.14323    0.08748  -1.637  0.10158
Anos8           -0.13880    0.09127  -1.521  0.12834
Anos10          -0.13544    0.08422  -1.608  0.10780
Depthr           0.04743    0.05786   0.820  0.41242
DeptIT           0.12084    0.05122   2.359  0.01831 *
Deptmanagement   0.09007    0.06066   1.485  0.13757
Deptmarketing    0.12460    0.05563   2.240  0.02513 *
Deptproduct_mng  0.15562    0.05477   2.842  0.00450 **
DeptRandD        0.07506    0.05635   1.332  0.18284
Deptsales        0.10792    0.04383   2.462  0.01382 *
Deptsupport      0.11232    0.04662   2.409  0.01599 *
Depttechnical    0.12809    0.04549   2.816  0.00487 **
---
```

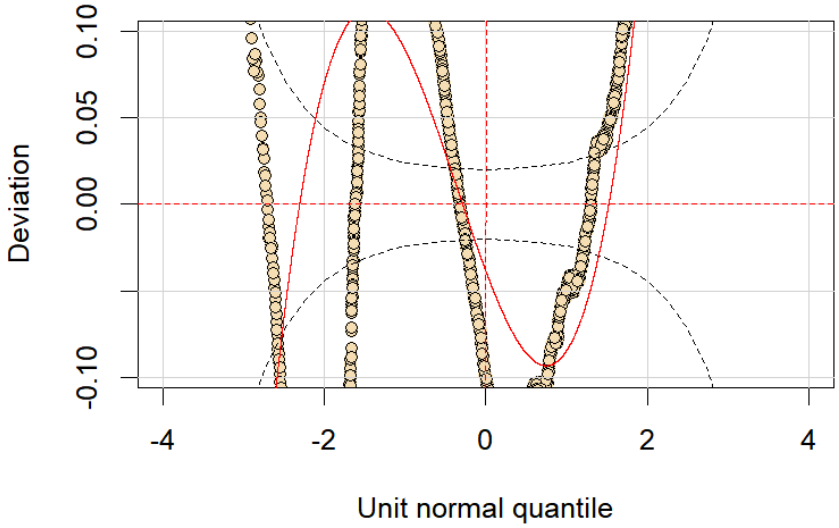
```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Sigma link function:  logit
Sigma Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.2952238  0.0488634   46.972  < 2e-16 ***
Aval        -0.2494892  0.0471778   -5.288 1.25e-07 ***
Proj3       -1.2467700  0.0263087  -47.390  < 2e-16 ***
Proj4       -1.2584003  0.0262134  -48.006  < 2e-16 ***
Proj5       -1.8143169  0.0290964  -62.355  < 2e-16 ***
Proj6       -4.2226245  0.0310873 -135.831  < 2e-16 ***
Proj7       -3.8623004  0.0507854  -76.051  < 2e-16 ***
Hora        -0.0010484  0.0001614   -6.494 8.64e-11 ***
Anos3       -0.1644752  0.0202652   -8.116 5.19e-16 ***
Anos4        0.0037232  0.0246342    0.151  0.87987
Anos5        0.1552486  0.0282641    5.493 4.02e-08 ***
Anos6        0.3369284  0.0359882    9.362  < 2e-16 ***
Anos7        0.0981560  0.0668032    1.469  0.14176
Anos8        0.0090098  0.0718178    0.125  0.90017
Anos10       0.1584949  0.0631298    2.511  0.01206 *
Acid1        0.1119669  0.0206556    5.421 6.03e-08 ***
Pgtolow     -0.1207065  0.0274221   -4.402 1.08e-05 ***
Pgtomedium  -0.0720707  0.0275526   -2.616  0.00891 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

------------------------------------------------------------------
Nu link function:  log
Nu Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -3.605e+01  1.431e-01 -251.978   <2e-16 ***
Proj3        3.490e+01  1.471e-01  237.179   <2e-16 ***
Proj4        3.536e+01  1.475e-01  239.815   <2e-16 ***
Proj5        3.712e+01  1.455e-01  255.142   <2e-16 ***
Proj6       -7.588e-03  1.638e-01   -0.046    0.963
Proj7       -5.674e-04  1.680e-01   -0.003    0.997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
-------------------------------------------------------------------
Tau link function:  log
Tau Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  3.652485   0.004068  897.832  < 2e-16 ***
Proj3       -1.812320   0.012921 -140.265  < 2e-16 ***
Proj4       -1.617584   0.013184 -122.695  < 2e-16 ***
Proj5       -3.148485   0.019083 -164.985  < 2e-16 ***
Proj6       -1.111983   0.006252 -177.862  < 2e-16 ***
Proj7       -1.065740   0.005340 -199.568  < 2e-16 ***
Acid1        0.023368   0.006925    3.375 0.000741 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

-------------------------------------------------------------------
NOTE: Additive smoothing terms exist in the formulas:
 i) Std. Error for smoothers are for the linear effect only.
ii) Std. Error for the linear terms may not be reliable.
-------------------------------------------------------------------
No. of observations in the fit:  14888
Degrees of Freedom for the fit:  62.53536
      Residual Deg. of Freedom:  14825.46
                      at cycle:  2683

Global Deviance:     -14682.18
            AIC:     -14557.11
            SBC:     -14081.32
*******************************************************************
```

Embora tenha sido seguido o protocolo descrito na seção de Metodologia, uma análise final do modelo, com avaliação gráfica por meio do *worm plot* (Figura 12), evidenciou inadequações significativas no ajuste, com desvios sistemáticos em praticamente toda a faixa dos quantis normais. Observa-se um padrão ondulado tanto nas caudas quanto no centro da distribuição, indicando que os resíduos não seguem a distribuição teórica esperada. A linha de ajuste (em vermelho) apresenta afastamentos consideráveis da linha de referência zero, e um número expressivo de pontos ultrapassa os limites das bandas de confiança, reforçando a má qualidade do ajuste. Vale destacar que a base de dados utilizada foi gerada de forma sintética, o que pode ter contribuído para esses resultados caso o processo de geração não tenha reproduzido adequadamente as características da

distribuição original.

Figura 12 – Worm plot dos resíduos do modelo com distribuição GB1



4 Considerações Finais

Embora tenha sido definido um protocolo de modelagem, reconhece-se que outras combinações de ajustes poderiam ter sido exploradas. A base de dados relacionada à felicidade apresentou um ajuste satisfatório, especialmente após a inclusão dos parâmetros de locação, escala e forma, o que contribuiu para a melhoria do modelo. A estrutura da base parecia bem organizada, o que favoreceu a obtenção de resultados consistentes. A interpretação dos resultados obtidos corrobora com as informações apresentadas no próprio Relatório de Felicidade disponibilizado pela organização responsável, evidenciando que variáveis relacionadas à qualidade de vida e fatores econômicos estão positivamente associadas à felicidade, sendo que países com melhores condições socioeconômicas apresentam pontuações mais elevadas.

Por outro lado, a base de dados referente à satisfação demonstrou comportamento atípico desde a análise descritiva inicial, com concentração de valores em faixas que não pareciam orgânicas. Mesmo seguindo um processo de modelagem semelhante ao adotado para a base anterior, o ajuste final não foi satisfatório.

As variáveis de quantidade de projetos em que o funcionário atua e anos de trabalho na empresa foram avaliadas de forma categórica, devido à baixa quantidade de níveis da categoria e à ausência de indícios de comportamento linear. No entanto, uma análise exploratória adicional, não apresentada neste trabalho, indicou que a utilização dessas variáveis em formato numérico poderia melhorar o ajuste do modelo. Fica, portanto, a sugestão de explorar essa abordagem em estudos futuros.

Como o objetivo principal do trabalho era demonstrar como diferentes metodologias de modelagem podem impactar significativamente os resultados, consideramos o resultado final satisfatório. Os achados reforçam a importância de uma escolha criteriosa da abordagem de modelagem, pois decisões metodológicas distintas podem levar a interpretações diferentes e, por vezes, conflitantes.

Referências

ATKINSON, A. C. The logitnormal distribution and its applications to modeling proportion data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Wiley, v. 34, n. 4, p. 338–343, 1985.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. *Journal of Multivariate Analysis*, v. 39, n. 1, p. 106–116, 1991.

BONAT, W. H.; RIBEIRO, P. J.; ZEVIANI, W. M. Regression models with responses on the unit interval: Specification, estimation and comparison. *Revista Brasieleira de Biomedicina*, v. 20, n. 1, p. 1–10, 2013.

CHARNET, R. et al. *Análise de Modelos de Regressão Linear com Aplicações*. São Paulo, Brasil: Unicamp, 2008. 368 p. ISBN 978-85-268-0780-8.

FERRARI, S. L. P.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004.

FERRARI, S. L. P.; LEMONTE, A. J.; CYSNEIROS, F. J. A. Hypothesis testing in heteroscedastic simplex regression models. *Journal of Statistical Planning and Inference*, v. 141, n. 1, p. 488–505, 2011.

GREENE, W. H. *Econometric Analysis*. Nova Jersey, Estados Unidos: Prentice Hall, 2002. 1024 p. (Prentice Hall). ISBN 978-0130661890.

HASTIE, T.; TIBSHIRANI, R. Generalized additive models. *Statistical Science*, Institute of Mathematical Statistics, v. 1, n. 3, p. 297–318, 1986.

HASTIE, T.; TIBSHIRANI, R. *Generalized Additive Models*. London: Chapman and Hall/CRC, 1990. (Monographs on Statistics and Applied Probability). ISBN 9780412343902.

HELLIWELL, J. F. et al. (Ed.). *World Happiness Report 2022*. New York: Sustainable Development Solutions Network, 2022.

MCCULLAGH, P.; NELDER, J. A. *Generalized Linear Models*. 2. ed. London: Chapman and Hall/CRC, 1989. (Monographs on Statistics and Applied Probability). ISBN 9780412317606.

MCDONALD, J. B.; XU, Y. J. A generalization of the beta distribution with applications. *Journal of Econometrics*, Elsevier, v. 66, n. 1-2, p. 133–152, 1995.

MENEZES, A. F. B.; FURRIEL, W. O. Modelos de regressão beta e simplex na análise do índice de desenvolvimento humano municipal de 2010. *Revista Brasileira de Biometria*, v. 37, n. 3, p. 394–408, 2019.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. Tempe, AZ: Wiley, 2012. 679 p. ISBN 978-0-470-54281-1.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society*, v. 135, n. 3, p. 370–384, 1972.

OSPINA, R.; FERRARI, S. L. Inflated beta distributions. *Statistical Papers*, Springer, v. 51, n. 1, p. 111–126, 2010.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Royal Society, v. 54, n. 3, p. 507–554, 2005.

STASINOPOULOS, M. D. et al. *Flexible Regression and Smoothing: Using GAMLSS in R*. Boca Raton, FL: Chapman and Hall/CRC, 2017. ISBN 9781498742051.

WEBER, R. J. *Regression Analysis: An Overview*. 2022. Kellogg School of Management, Northwestern University. Aulas e material disponível publicamente no site da Kellogg. Disponível em: <https://www.kellogg.northwestern.edu/faculty/weber/jhu/statistics/regression.htm>.