**Universidade de Brasília**
**Departamento de Estatística**

# Análise das despesas administrativas de fundos de pensão por meio de regressão quantílica linear longitudinal

**Leonardo Almeida de Magalhães**

**Brasília**
**2023**

**Leonardo Almeida de Magalhães**

# Análise das despesas administrativas de fundos de pensão por meio de regressão quantílica linear longitudinal

Orientador: Prof. Eduardo Yoshio Nakano

Relatório final de Trabalho de Conclusão de Curso apresentado para o Departamento de Estatística da Universidade de Brasília, como requisito para obtenção do grau de Bacharel em Estatística.

**Brasília**
**2023**

# Resumo

O nível de despesas administrativas incorridas pelos fundos de pensão tem o potencial de impactar diretamente nos benefícios previdenciários pagos por esses aos participantes. A fim de compreender os fatores que possam explicar os diferentes níveis de despesas e eventuais fontes de ineficiência na gestão administrativa, aplicou-se análise de regressão quantílica para medidas repetidas em modelos lineares com intercepto aleatório. Os resultados indicaram bom ajustamento do modelo e o uso de transformação logarítmica das variáveis resposta e explicativas quantitativas. Efeitos significativos foram observados para as covariáveis: valor do ativo, tipo de patrocínio do fundo de pensão (se realizado por entes públicos ou privados), tempo de existência, quantidade de planos de benefícios administrados e percentual de aplicação em imóveis.

Palavras-chaves: fundo de pensão, despesas administrativas, modelos lineares mistos, regressão linear quantílica longitudinal

# Lista de Tabelas

# Lista de Figuras

# Sumário

# 1 Introdução

Os fundos de pensão, denominados Entidades Fechadas de Previdência Complementar (EFPC) na legislação brasileira, são pessoas jurídicas sem fins lucrativos que têm como principal objetivo pagar benefícios de caráter previdenciário para os seus "participantes e assistidos", os empregados das entidades que as institui, as chamadas "patrocinadoras".(BRASIL, 2001).

Para atingir objetivo de pagar benefícios previdenciários, os participantes e patrocinadores realizam contribuições periódicas para as EFPC, que por sua vez realizam investimentos no mercado financeiro com o intuito de capitalizar os recursos para o pagamento futuro desses benefícios. Nos casos em que as regras que definem a forma de pagamento dos benefícios são afetadas por eventos incertos, por exemplo, em relação à duração do pagamento ou à própria ocorrência, são aplicados cálculos atuariais para realizar as estimativas dos recursos necessários para fazer frente às obrigações projetadas. Para executar essas atividades os fundos de pensão incorrem em gastos administrativos que envolvem a manutenção de infraestruturas físicas e tecnológicas e a remuneração de pessoal especializado, dentre outras despesas.

De acordo com publicação periódica do órgão supervisor dos fundos de pensão, a Superintendência Nacional de Previdência Complementar - Previc, o valor das despesas administrativas é influenciado diretamente por características das EFPC, que podem ser bastante diversas, tais quais tempo de funcionamento e volume de recursos geridos (PREVIC, 2021).

Algumas dessas características são reconhecidas por apresentarem associação estatística entre o valor de despesa administrativa e a característica observada. Em alguns casos essa associação teria uma fundamentação econômica, como é o caso do porte (escala) da EFPC: fundos de pensão que gerenciam maior volume de recursos tendem a ter maiores despesas, tudo mais constante (DICK; POMORSKI, 2010). Entretanto é possível hipotetizar que nem todas as características que apresentem associação estatística com as despesas teriam relação de causalidade sob o que se esperaria pela lógica econômica.

Esse contexto, em que diversas características podem "justificar" valores esperados diferenciados de despesas administrativas das EFPC, impossibilita comparação direta entre as EFPC com relação ao grau de eficiência na gestão administrativa. Contudo, avalia-se que a aplicação de modelos estatísticos, em especial aqueles cuja relação funcional seja de natureza aditiva ou linear, pode auxiliar nesse processo, ao permitir decompor

o valor predito pelos efeitos de cada covariável.

Entende-se que a aplicação de modelo estatístico para ampliar a compreensão sobre o comportamento das despesas administrativas teria utilidade para os órgãos de governança das EFPC (Conselhos Deliberativos e Fiscais, Diretoria e comitês internos), possibilitando evidenciar a necessidade de ajustes na gestão administrativa do fundo de pensão com maiores indícios de ineficiência com vistas a reduzir os gastos, o que no limite se reverteria em maiores benefícios para os participantes e assistidos.

Por outra perspectiva, acredita-se que o modelo também pode ser aplicado pela Previc na supervisão dos fundos de pensão. A identificação das EFPC menos eficientes permitiria ações de monitoramento e fiscalização no sentido de orientá-las à redução das despesas avaliadas, por meio do modelo adotado, como mais elevadas que as "necessárias" para permitir uma adequada administração, já consideradas as particularidades de cada EFPC. Também se avalia haver potencial na identificação de eventuais imposições regulatórias cujos custos implicados para as EFPC possam superar os benefícios percebidos por esse órgão.

O objetivo deste trabalho é, portanto, modelar os gastos administrativos dos fundos de pensão brasileiros por meio de relação funcional linear. Todavia, devido à grande heterogeneidade dos fundos de pensão existentes e presença de valores atípicos de despesas o pressuposto de normalidade dos modelos de lineares clássicos podem não ser adequado, bem como os métodos de estimação de parâmetros por mínimos quadrados. Considerando essas questões, neste trabalho propõe-se o uso de regressão quantílica linear como abordagem alternativa a fim de se obter um melhor ajustamento e redução de vieses nos coeficientes estimados.

A seção 2 traz breve revisão do arcabouço teórico utilizado para o desenvolvimento deste trabalho, na subseção 2.1, fazem-se considerações sobre a estrutura funcional do modelo na subseção, enquanto a subseção 2.2 apresenta-se o modelo de regressão quantílica aplicado aos dados coletados das EFPC. Em seguida abordam-se aspectos metodológicos do trabalho na seção 3, para então serem apresentados os resultados, seção 4, em que se faz primeiramente uma análise exploratória (subseção 4.1) seguida pela apresentação do modelo final proposto (subseção 4.2). As conclusões do trabalho são apresentadas ao final, na seção 5.

# 2 Referencial Teórico

## 2.1 Estrutura funcional do modelo

De acordo com Goldfeld e Quandt (1970), a função de produção de Cobb-Douglas é frequentemente utilizada para modelar variáveis econômicas, sendo originalmente proposta para considerar a produção como variável resposta $v_i$, $i = 1, \ldots, n$. Considerando o vetor de variáveis explicativas $\mathbf{z}_i = (z_{1,i}, z_{2,i}, \ldots, z_{p,i})$, o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ associado a $\mathbf{z}_i$ e um componente de erro aleatório multiplicativo $e^{\varepsilon_i}$, a referida função é descrita como

$$v_i = e^{\beta_0} z_{1,i}^{\beta_1} z_{2,i}^{\beta_2} \ldots z_{p,i}^{\beta_p} e^{\varepsilon_i}. \tag{1}$$

A produção, que no caso de EFPC são os serviços prestados para os participantes, está intimamente relacionada com o montante de despesas, portanto é natural a suposição de que essas possam ser modeladas por meio de relações funcionais multiplicativas conforme a equação 1. Com vistas a linearizar essa relação de tal forma a permitir que modelos aditivos possam ser aplicados e, por consequência, interpretar os $\beta_j$, $j = 0, 1, \ldots, p$, em termos das variações absolutas das covariáveis, é possível aplicar as transformações logarítmicas $y_{j,i} = log(v_{j,i})$ e $x_{j,i} = log(z_{j,i})$ , fazendo com que o modelo assuma a forma

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} + \varepsilon_i. \tag{2}$$

Ocorre que em análises econométricas é bastante comum que dados medidos sobre as unidades observacionais possam ser referenciados tanto sob a perspectiva da unidade em si quanto no momento em que é observado. Modelos dessa natureza são chamados de longitudinais ou de dados em painel. Conforme Hsiao (1986) o valor da variável resposta do $i$-ésimo indivíduo para o $t$-ésimo tempo ($y_{i,t}$) pode ser descrito pelo o modelo linear irrestrito

$$y_{i,t} = \beta_{0,i,t} + \sum_{j=1}^{p} \beta_{j,i,t} x_{j,i,t} + \varepsilon_{i,t}. \tag{3}$$

É notável na equação 3 a grande quantidade de parâmetros, contudo no processo

de modelagem é comum a busca por estruturas funcionais mais simples. Bozdogan (1987) destaca inconveniências no uso de modelos excessivamente complexos como aumento de custos de medição e riscos de superparametrização. Nesse sentido, modelos lineares mais parcimoniosos podem ser obtidos quando se supõe invariância dos parâmetros com relação ao indivíduo, ao tempo ou a ambos. Neste último caso, em que $\beta_{j,i,t} = \beta_j$, $j = 0, 1, \ldots, p$, obtém-se a expressão do modelo de regressão linear simples ou *pooled* tal qual a representada pela equação 2 (HSIAO, 1986).

Hsiao (1986) descreve uma aplicação da função de Cobb-Douglas para análise da produção em contextos de dados em painel. Destaca, em particular, que a suposição de que o coeficiente de intercepto seja invariante em relação aos indivíduos e ao tempo seria passível de críticas por ignorar variáveis que poderiam refletir capacidade de gestão e outras diferenças de técnicas entre firmas ou variáveis que afetem a produtividade de todas as firmas mas que variem com o tempo, como por exemplo condições climáticas para produção agrícola.

De forma mais geral, talvez a maior ressalva ao se pressupor a aplicação de um modelo *pooled* para dados longitudinais, se refira ao fato de que essa simplificação pressupõe que as covariáveis $x_{i,t}$ são independentes, o que em regra não se sustenta pois observações de um mesmo indivíduo tendem a ser naturalmente dependentes e isso deve ser levado em consideração para evitar vieses nas estimativas dos parâmetros (MARINO; FARCOMENI, 2015).

Abordagem tradicional para análise de dados longitudinais pressupõe os componentes de erro $\varepsilon_{i,t}$ independentes e normais com média zero e variância constante $\sigma^2$. Todavia, não raramente a premissa de normalidade ou mesmo distribuição homogênea para os termos de erros é inapropriada para dados reais nos mais diversos campos de estudos. A regressão quantílica relativiza algumas dessas pressuposições e a fim de introduzir essa técnica de análise faz-se primeiramente breve revisão da abordagem transversal, considerando as observações independentes, para na seção seguinte incluir a suposição de dependência entre covariáveis medidas sobre um mesmo indivíduo, por meio de modelo com intercepto aleatório.

## 2.2 Regressão quantílica

### 2.2.1 Regressão quantílica linear para observações independentes

Ao passo que as técnicas tradicionais de regressão linear ou mesmo de análise de dados em painel focam na função de regressão, ou seja, no valor esperado da variável resposta, $Y$, condicionada a valores do conjunto de variáveis explicativas $\boldsymbol{X}$ a regressão quantílica estende essa abordagem no sentido de permitir o estudo diretamente (sem a necessidade, a priori, de recorrer a pressuposições para a distribuição do termo de erro do modelo) da distribuição de $Y$ condicionada a $\boldsymbol{X}$ (DAVINO MARILENA FURNO, 2014).

A regressão quantílica foi introduzida por Koenker e Jr (1978), nesse trabalho destacam a utilidade e a maior robustez de métodos de minimização baseados em desvios absolutos ao lidar com *outliers* ou dados cuja confiabilidade possa ser reduzida. Enfatizam ainda que os métodos por mínimos quadrados ordinários são muito sensíveis a uma modesta "contaminação" por *outliers* tornando o método um preditor ruim especialmente para distribuições com caldas pesadas. Nesse sentido, por serem tradicionalmente marcados pela presença de valores extremos, Huang Hanze Zhang e He (2011) destacam a relevância da regressão quantílica para modelagem de dados econômicos e financeiros, categoria na qual se inserem os dados que se pretende modelar.

Enquanto a estimação dos parâmetros do modelo linear transversal por meio do método de mínimos quadrados ordinários baseia-se na estimação da esperança condicional $\mu(Y \mid \mathbf{X} = \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$ por meio da equação

$$\hat{\mu}(y_i \mid \mathbf{X} = \mathbf{x}_i) = \underset{\mu}{\operatorname{argmin}}\{E\left[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right]\}. \tag{4}$$

Na abordagem da regressão quantílica o objetivo é caracterizar a distribuição de $y_i \mid \mathbf{x}_i$ por meio dos quantis, os quais correspondem à inversa da função de distribuição. Nessa técnica de análise, para a estimação de parâmetros, a função quadrática do método de mínimos quadrados ordinários é substituída pela função de perda absoluta assimétrica $\rho_\tau(u) = u[\tau - \mathbb{I}(u < 0)]$, em que $\mathbb{I}(\cdot)$ é a função indicadora, e dessa forma a função de perda associa pesos $\tau \in (0, 1)$ e $(\tau - 1)$ a desvios positivos e negativos, respectivamente. Portanto, supondo-se o modelo linear para um dado quantil

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)} + \varepsilon_i^{(\tau)}, \tag{5}$$

em que o quantil $\tau$ da distribuição do termo de erro condicionado às covariáveis é nulo ($Q_\tau\left(\varepsilon_i \mid \boldsymbol{\beta}, \mathbf{x}_i\right) = 0$), tem-se que a estimativa do quantil $Q_\tau\left(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i\right) = F_\tau^{-1}(y_i \mid \boldsymbol{\beta}, \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)}$ é obtida pela otimização

$$\hat{Q}_\tau\left(y_{it} \mid u_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right) = \underset{Q}{\operatorname{argmin}} E\{\left[\rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)})\right]\}. \tag{6}$$

Por sua vez, os coeficientes do modelo linear para o quantil $\tau$ condicionado aos valores das covariáveis observadas pode ser estimados pela função

$$\hat{\boldsymbol{\beta}}^{(\tau)} = \arg\min_{\boldsymbol{\beta}^{(\tau)}} \sum_{i:y_i \geq \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)}} \tau \left|y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)}\right| + \sum_{i:y_i < \mathbf{x}_i'\beta^{(\tau)}} (1-\tau) \left|y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)}\right|. \tag{7}$$

Para se solucionar esse problema de minimização dado pela equação 7, Geraci e Bottai (2007) propõem a suposição auxiliar de que $\varepsilon_i^{(\tau)}$ tem distribuição de Laplace assimétrica (AL) com parâmetro de posição nulo. Uma variável $W \sim AL(\mu, \sigma, \tau)$, quando sua função de densidade é descrita por

$$f_W(w \mid \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_\tau(w-\mu)\right\}, \quad -\infty < w < \infty, \tag{8}$$

em que $\rho_\tau(\cdot)$ é a função de perda assimétrica definida anteriormente, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_+$ e $0 < \tau < 1$ são os parâmetros de posição, escala e assimetria, respectivamente.

É possível verificar que o parâmetro de posição $\mu$ é o $\tau$-ésimo quantil de $W$, ou seja, $\mathrm{P}(W \leq \mu) = \tau$. Dessa forma a premissa de que $\varepsilon_i^{(\tau)} \sim \mathrm{AL}(0, \sigma^{(\tau)}, \tau)$ no modelo linear transversal (equação 5) permite reformular o problema de otimização da regressão quantílica para uma abordagem baseada em máxima verossimilhança. Uma vez que minimizar a equação 7 equivale a maximizar a função de verossimilhança

$$L^{(\tau)}(\boldsymbol{\beta}, \sigma, \tau) = \left[\frac{\tau(1-\tau)}{\sigma^{(\tau)}}\right]^n \exp\left\{-\frac{1}{\sigma^{(\tau)}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}^{(\tau)})\right\}. \tag{9}$$

Adicionalmente a assunção dessa distribuição permite estender o modelo para o contexto de interesse, em que há dependência entre as observações, como na abordagem longitudinal (MARINO; FARCOMENI, 2015).

### 2.2.2 Regressão quantílica longitudinal com intercepto aleatório

Geraci e Bottai (2007) propuseram modelo de regressão linear longitudinal para intercepto aleatório recorrendo à suposição de que a variável resposta, condicionada a um dado valor do intercepto aleatório, tem distribuição de Laplace assimétrica a fim de permitir a aplicação do ferramental desenvolvido para a teoria de estimação de por máxima verossimilhança.

A equação de regressão do modelo linear longitudinal referido para um dado quantil $\tau$ apresenta a seguinte forma

$$y_{i,t} = u_i + \mathbf{x}'_{i,t}\boldsymbol{\beta}^{(\tau)} + \varepsilon^{(\tau)}_{i,t}, \tag{10}$$

em que $y_{i,t}$ é a variável resposta do $i$-ésimo indivíduo ($i = 1, \ldots, n$) no $t$-ésimo tempo ($t = 1, \ldots, n_i$), $u_i$ é variável aleatória que depende do indivíduo, correspondente ao intercepto aleatório, e $\varepsilon^{(\tau)}_{i,t}$ é o termo, para o qual o quantil condicional $\tau$ é nulo, ou seja, $Q_\tau\left(\varepsilon_{i,t} \mid u_i, \boldsymbol{\beta}, \mathbf{x}_{i,t}\right) = 0$. Os $u_i$ são mutuamente independentes e identicamente distribuídos de acordo com uma função de densidade $f\left(u_i \mid \varphi^{(\tau)}\right)$ definida por um parâmetro $\varphi^{(\tau)}$ dependente de $\tau$. Também se assume que $\varepsilon_{i,t}$ são independentes e que $u_i$ e $\varepsilon_{i,t}$ são independentes entre si.

Assume-se, adicionalmente, que $y_{i,t}$ condicionado a $u_i$ tem distribuição de Laplace assimétrica com parâmetro de posição igual ao preditor linear $u_i + \mathbf{x}'_{i,t}\boldsymbol{\beta}^{(\tau)}$ e são independentes entre si.

Assim dada a propriedade discutida anteriormente para a distribuição de Laplace assimétrica, o quantil $\tau$ da distribuição de $y_{i,t} \mid u_i$ é igual ao preditor linear do modelo descrito pela equação 10, ou seja,

$$Q_\tau\left(y_{i,t} \mid u_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right) = u_i + \mathbf{x}'_{i,t}\boldsymbol{\beta}^{(\tau)}. \tag{11}$$

Fazendo, $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,n_i})$ e $f\left(\mathbf{y}_i \mid \beta, u_i, \sigma\right) = \prod_{t=1}^{n_i} f\left(y_{i,t} \mid \beta, u_i, \sigma\right)$ a densidade do $i$-ésimo indivíduo condicionado ao intercepto aleatório $u_i$. A densidade conjunta de $(\mathbf{y}_i, u_i)$, para $i = 1, \ldots, n$, é

$$f^{(\tau)}\left(\mathbf{y}_i, u_i \mid \boldsymbol{\beta}, \sigma, \varphi\right) = f\left(\mathbf{y}_i \mid \boldsymbol{\beta}^{(\tau)}, u_i, \sigma^{(\tau)}\right) f\left(u_i \mid \varphi^{(\tau)}\right). \tag{12}$$

Com base nessa distribuição conjunta de $(\mathbf{y}_i, u_i)$ é possível estimar os parâmetros $(\beta, \sigma, \varphi)$

por meio de estimação de máxima verossimilhança marginal considerando que

$$L_i^{(\tau)}\left(\boldsymbol{\beta}, \sigma, \varphi \mid \mathbf{y}_i\right) = \int_{-\infty}^{\infty} f\left(\mathbf{y}_i \mid \boldsymbol{\beta}^{(\tau)}, u_i, \sigma^{(\tau)}\right) f\left(u_i \mid \varphi^{(\tau)}\right) du_i, \tag{13}$$

A solução dessa integral em regra não possui forma fechada, portanto sendo necessário lançar mão de métodos numéricos. Geraci e Bottai (2007) propuseram o uso do algoritmo de Monte Carlo EM, que foi aplicado para modelos lineares mistos generalizados. Contudo, em função de o método ser computacionalmente ineficiente, os mesmos autores em trabaho posterior sugeriram a aplicação de métodos de quadratura Gaussiana (GERACI; BOTTAI, 2014). Para esses procedimentos são utilizadas as suposições de que $u_i \sim N(0, \varphi^{(\tau)})$ ou que $u_i \sim \text{AL}(0, \varphi^{(\tau)}, \frac{1}{2})$, entretanto destacam que é possível aplicação de diversas outras distribuições simétricas ou não.

# 3 Metodologia

As informações referentes à variável resposta e às explicativas foram originadas de bancos de dados administrados pela Previc, os quais advém de demonstrativos contábeis, de investimentos, atuariais e de dados cadastrais e são informados remotamente pelas EFPC via sistemas informatizados. A amostra utilizada compreende dados de 313 EFPC referentes aos exercícios fiscais de 2011 a 2021, perfazendo um total de 2981 observações. Os dados não são balanceados uma vez que as EFPC podem entrar ou sair da amostra ao longo do período considerado.

Com vistas a se definir uma relação de potenciais variáveis explicativas realizou-se pesquisa prévia com servidores públicos que trabalham na Previc e em seguida realizou-se refinamento por meio de análise descritiva bivariada dos dados. A análise exploratória também teve o propósito de verificar dependência intra-indivíduo para se avaliar a conveniência de aplicação de modelo que leve em consideração a repetição de medidas.

As análises descritivas e inferenciais foram realizadas por meio do software R Core Team (2021) e para aplicação do modelo de regressão quantílica longitudinal foi utilizado o pacote *lqmm* (GERACI, 2014). O processo de seleção de variáveis explicativas se amparou na avaliação da significância dos coeficientes e na minimização do critério de informação de Akaike (AIC).

# 4 Resultados

## 4.1 Descrição dos dados e análise exploratória

A variável resposta e potenciais variáveis explicativas a serem consideradas na análise, com as respectivas descrições, estão dispostas na Tabela 1.

Tabela 1: Descrição da variável resposta e potenciais variáveis explicativas

| Nome | Codificação | Descrição |
|---|---|---|
| Despesas administrativas **(variável resposta)** | DESP | Montante financeiro (em reais) despendido pela EFPC no período de um exercício fiscal |
| Ativo | ATIVO | Saldo final do ativo (em reais) ao final do exercício fiscal |
| Quantidade de planos | QTPLAN | Quantidade de planos administrados pela EFPC ao final do exercício fiscal |
| Idade | IDADE | Tempo em anos decorrido entre a criação e o final do exercício fiscal |
| Proporção de provisões BD | PROVBD | Razão entre as provisões matemáticas de benefício definido e as provisões matemáticas totais ao final do exercício fiscal |
| Proporção de ações | ACAO | Razão entre o valor aplicado em ações e o ativo ao final do exercício fiscal |
| Proporção de operações com participantes | OPPART | Razão entre o valor aplicado em operações com participantes e o ativo ao final do exercício fiscal |
| Proporção de FIP e FIDC | FIPFIDC | Razão entre o valor aplicado em operações com participantes e o ativo ao final do exercício fiscal |
| Proporção de títulos privados | TITPRIV | Razão entre o valor aplicado em títulos privados e o ativo ao final do exercício fiscal |
| Proporção de imóveis | IMÓVEL | Razão entre o valor aplicado em imóveis e o ativo ao final do exercício fiscal |
| Tipo de patrocínio | PATR | Natureza jurídica predominante dos patrocinadores dos planos administrados pela EFPC, assume o valor um, se pública, e zero, se privada |

As variáveis ACAO, FIPFIDC, IMOVEL, OPPART e TITPRIV, na escala original (valores monetários), correspondem a classes de investimentos e compõem o ativo das EFPC, portanto a associação substancial entre essas variáveis e o ATIVO é naturalmente esperada. De forma semelhante, as provisões matemáticas (PROVBD na escala original) compõem o passivo das EFPC, que por sua vez equivale ao ativo. Assim, para se reduzir efeitos de multicolinearidade essas variáveis foram reescaladas e avaliadas como proporção do ATIVO.

Verificada a presença de valores atípicos de despesa com frequência acima do padrão do restante dos dados para os casos em que ATIVO é inferior a dez milhões de reais ou nos quais a EFPC tem menos de seis meses de existência (IDADE $< 0,5$), optou-se por excluir essas observações. Dessa forma o modelo final proposto não tem como intuito explicar ou prever a distribuição condicional de DESP para observações que se enquadrem nos critérios mencionados.

Análises descritivas bivariadas foram realizadas para identificar potenciais padrões de relacionamento entre as variáveis. Nesse sentido nota-se uma flagrante associação positiva entre o ativo das EFPC e a despesa administrativa, conforme se observa pela Figura 1, embora aparentemente não seja linear, revelando ganhos de escala, conforme sugerido pelos estudos de Dick e Pomorski (2010), ou seja, a razão entre a variação de DESP e de ATIVO decresce, embora sempre positiva, a medida que ATIVO cresce. Essa hipótese de ganhos de escala é compatível com a suposição de aplicação de um modelo multiplicativo do tipo Cobb-Douglas, no qual o expoente associado à covariável ATIVO seria positivo e inferior a 1 (um).

Conforme argumentado, a transformação das variáveis pela função logarítmica tem como efeito a linearização do padrão de associação, o que é notado visualmente a partir da análise da Figura 3 e pelo incremento do coeficiente de correlação linear de Pearson (de 0,83 para 0,86). É possível especular ainda que esse aumento da correlação não tenha sido superior em função da existência de poucas EFPC com ativos mais elevados, pois, caso houvesse, a caracterização associação curvilínea poderia ser ainda mais evidente, supondo-se verdadeira a hipótese de ganhos de escala.

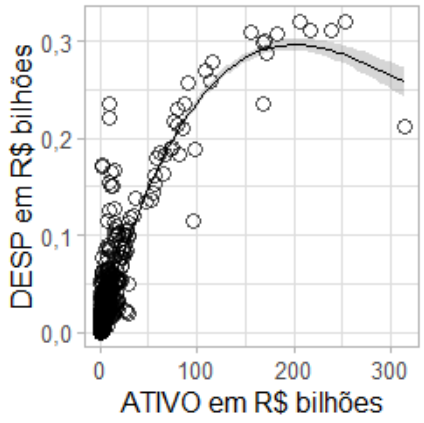


Figura 1: Gráfico de dispersão entre ATIVO e DESP, linha de tencência ajustada para as duas variáveis a partir de Modelo Aditivo Generalizado (GAM) com alisamento por *B-spline* sugere associação não linear (MARX; EILERS, 1998)

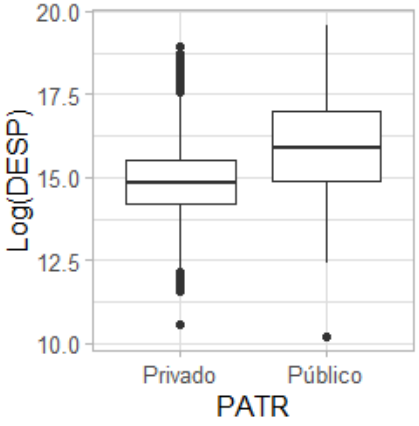


Figura 2: Diagrama de caixas do logarítmo de DESP para cada um dos níveis da variável PATR

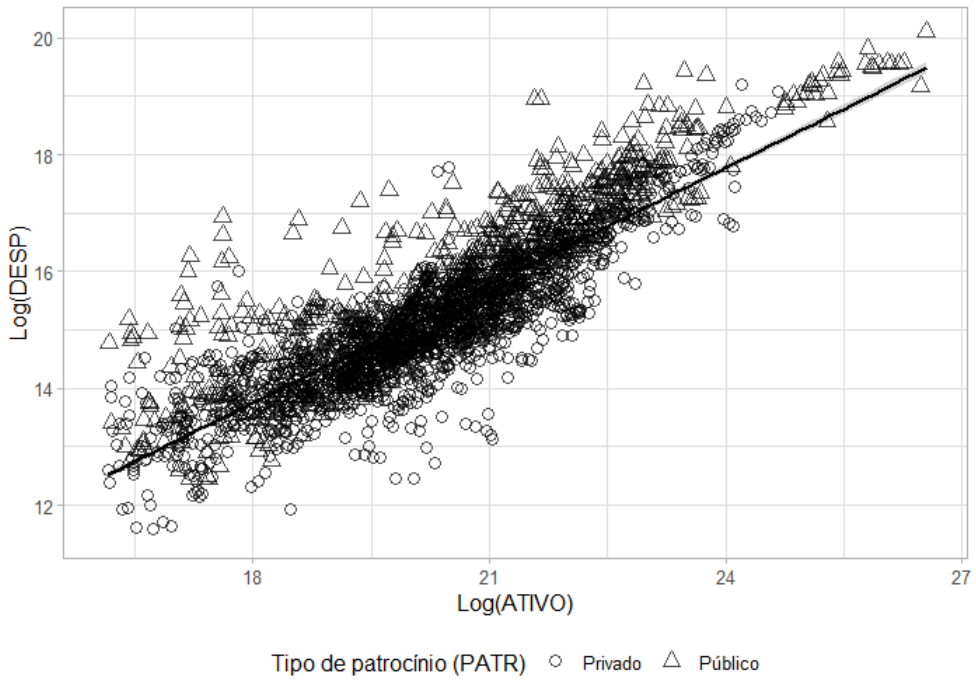


Figura 3: Gráfico de dispersão entre o logaritmo do ativo e das despesas administrativas com o ajustamento de reta de regressão linear simples e identificação dos níveis da variável PATR

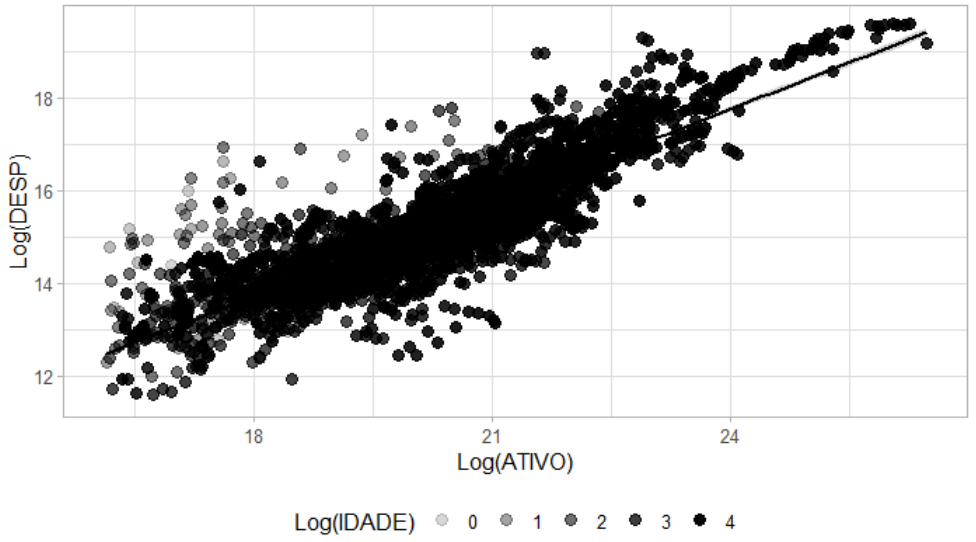


Figura 4: Gráfico de dispersão entre o logaritmo do ativo e das despesas administrativas com o ajustamento de reta de regressão linear simples. O grau de clareamento dos pontos representa o logaritmo de IDADE

Um efeito adicional da transformação logarítmica é o de potencial redução da heterocedasticidade (SCHMIDT; GERMANO; MILANI, 2019). É possível notar a partir da Figura 1 um aumento da dispersão da despesa para valores mais elevados do ativo,

ao passo que pela Figura 3 verifica-se um padrão mais homogêneo. Embora neste caso perceba-se maior dispersão para valores menores do ativo, aparentemente essa variação é menor que a notada na Figura 1. De fato, quando aplicado ao modelo de regressão linear simples, a estatística do teste de Breuch-Pagan (BREUSCH; PAGAN, 1979) reduziu de 538 para 24, indicando menor heterocedasticidade. Novamente é possível conjecturar que a presença de mais observações que apresentassem valor do ATIVO elevado evidenciaria com mais clareza essa análise. Haja vista a premissa do modelo aplicado de distribuição idêntica do termo de erro a diminuição da heterocedasticidade é conveniente.

A Figura 2 revela um aparente efeito significativo do tipo de patrocínio (PATR) sobre DESP. As distâncias entre os quantis correspondentes nos diagramas fornecem indícios de diferença na distribuição do logaritmo de DESP condicionado ao valor da variável PATR. O potencial explicativo dessa variável também pode ser observado pela figura 3 que revela uma concentração de EFPC cujo patrocínio predominante é público acima da reta de regressão ajustada para os logaritmos do ATIVO e da DESP, o que sugere que após considerado o efeito de ATIVO, o efeito de PATR também seria significativo.

Embora mais sutil, análise semelhante a essa realizada para a relação entre ATIVO e a variável resposta pode ser realizada para as demais variáveis quantitativas incluídas na análise com potencial de explicação das despesas. Com base nessas análises, na aparente redução de heterocedasticidade (ao menos para o efeito do ATIVO, variável explicativa que se mostra como a mais significativa) e na suposição de aplicabilidade de modelo Cobb-Douglas, todas as covariáveis quantitativas foram log-transformadas. A interpretabilidade dos parâmetros estimados pode se somar às justificativas para a essa transformação, uma vez que, a presença dessas na escala original juntamente com as transformadas dificultaria a compreensão geral dos efeitos e do significado prático do modelo sob a perspectiva econômico-financeira. Será utilizado o prefixo “LOG” quando em referência à variável transformada pela função logarítimica.

A Figura 5 apresenta-se um quadro de correlações entre variáveis quantitativas log-transformadas.

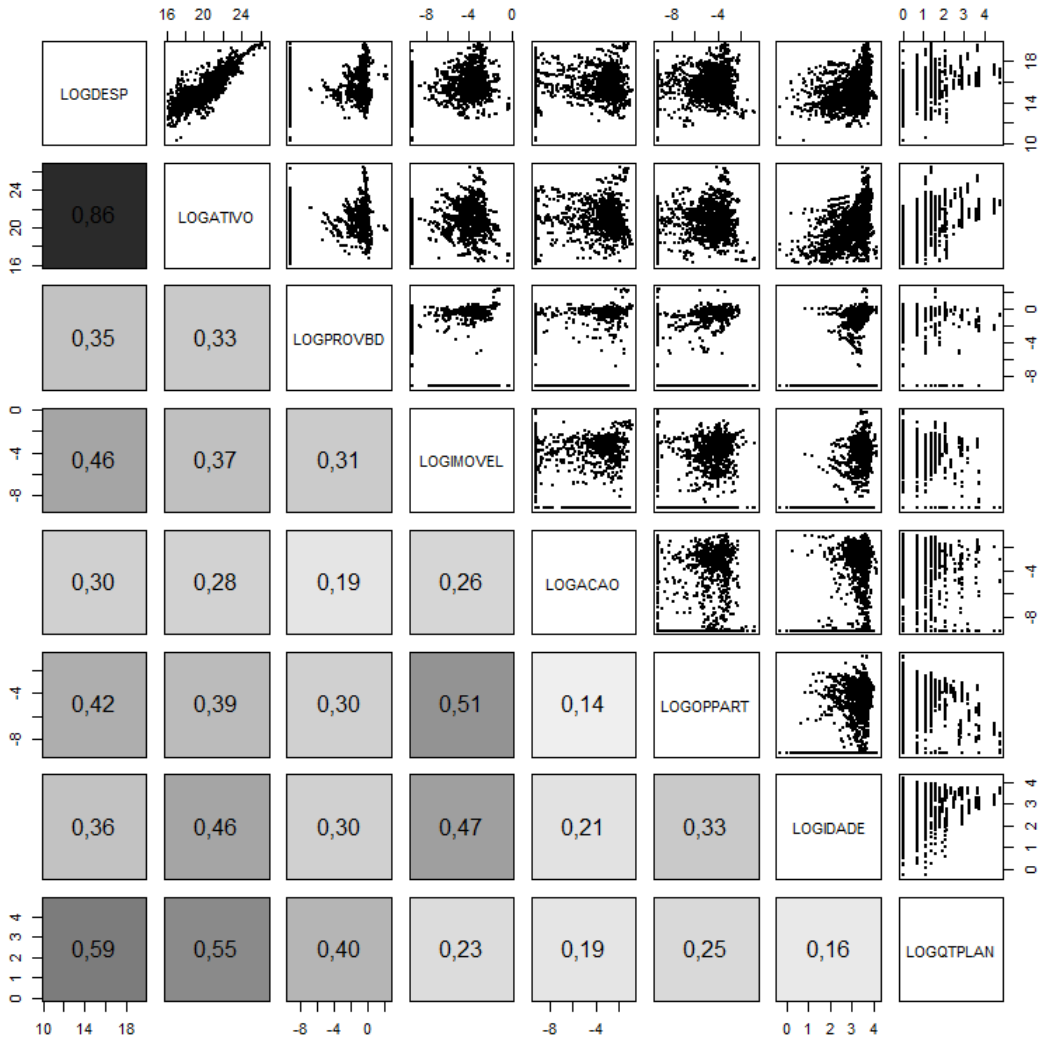


Figura 5: Gráficos de dispersões e correlações estatisticamente significativas (ao nível de significância de 0,05) entre o logaritmo das variáveis quantitativas que apresentaram maiores correlações com o logaritmo das despesas administrativas (DESP)

Efeito particular se nota em relação à variável LOGIDADE. Embora a correlação dessa com LOGDESP seja positiva, esse resultado possivelmente se deve à correlação relativamente alta entre LOGIDADE e LOGATIVO. Analisando-se a figura 4 nota-se uma uma maior frequência de observações em que LOGIDADE é menor que 2 acima da reta de regressão, o que indica que após o efeito de LOGATIVO a LOGIDADE teria um efeito potencialmente negativo sobre a variável resposta. Embora menos extremo, fenômeno semelhante pode ser observado para as variáveis LOGQTPLAN e LOGPROVBD. É plausível, portanto, que o aparente efeito substancial dessas variáveis sobre a variável resposta, seja inflado pela correlação dessa com LOGATIVO.

Previc (2021) destaca a heterogeneidade do sistema de previdência complementar

fechado, e diferenças entre estruturas das EFPC e dos planos de benefício, que refletiria em diferentes modelos de negócio e por consequência custos de administração. Essa variabilidade nos gastos administrativos pode ser percebida pela grande presença de valores extremos (*outliers*), que por meio dessa análise descritiva preliminar não seriam facilmente explicados pelas covariáveis, conforme se observa tanto pela Figura 2 quanto pelas Figuras 3 e 5. Esse aspecto representa um fator dificultador adicional para o processo de modelagem e uma das justificativas para adoção de modelo baseado em regressão quantílica, cuja função de perda penaliza mais moderadamente valores extremos que métodos baseados em mínimos quadrados. Ainda com relação a esse ponto, espera-se que a aplicação de modelo baseado em interceptos que variam por indivíduo, implique num menor reflexo desses *outliers* nos resíduos do modelo.

Para fins de ilustração do efeito de dependência intra-indivíduo, aplicou-se modelo de regressão linear múltipla transversal (*cross-sectional*), considerando as mesmas covariáveis do modelo final deste estudo, o qual será detalhado mais adiante e que incluem as variáveis LOGATIVO, PATR, LOGQTPLAN, LOGIDADE e LOGIMOVEL. Os resíduos do modelo foram dispostos na Figura 7, agrupados por EFPC, e revelam aparente dependência entre as observações para um mesmo indivíduo. Por outro lado a Figura 6 não fornece indícios quanto à existência de efeito temporal comum a todos os indivíduos.

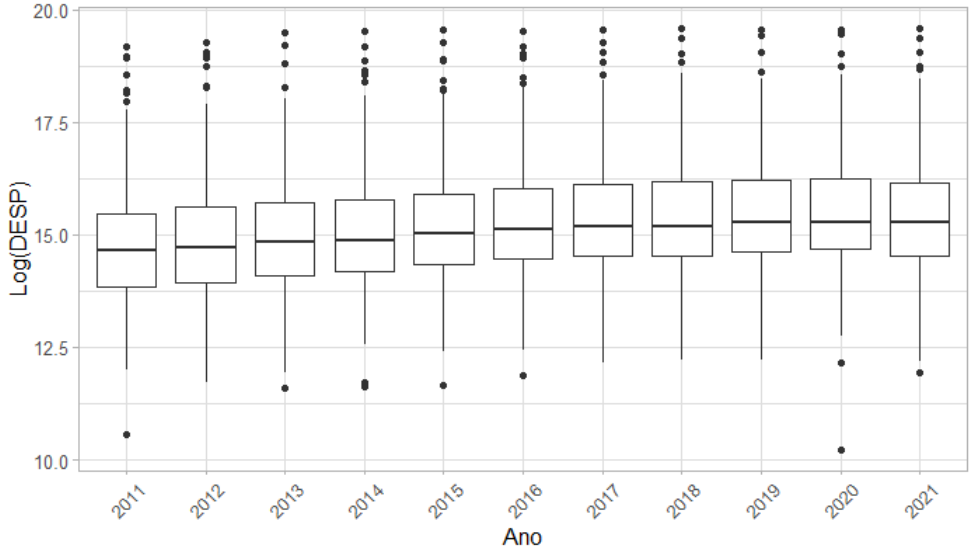


Figura 6: Diagrama de caixas do logarítmo de DESP para cada um dos anos de coleta dos dados

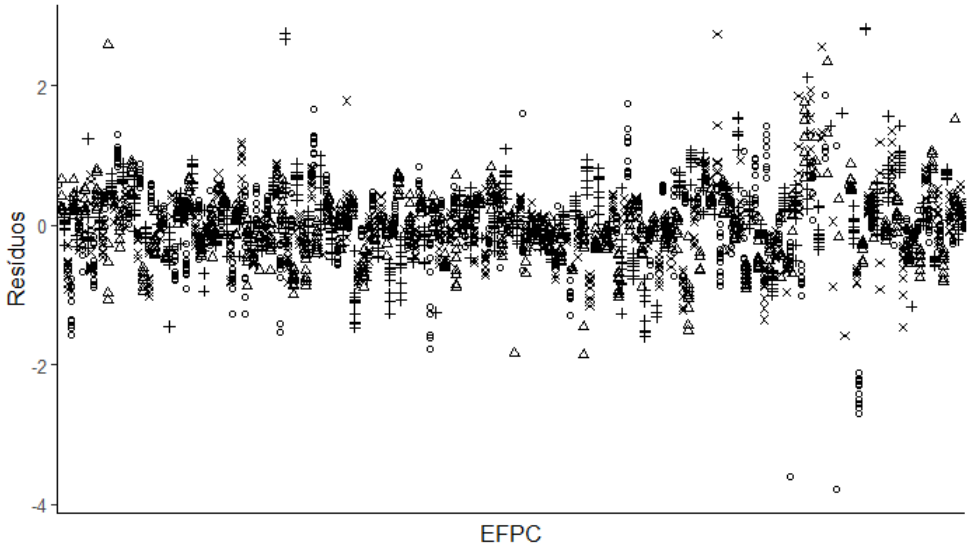


Figura 7: Gráfico de dispersão dos resíduos modelo de regressão múltipla para dados independentes considerando as covariáveis do modelo final (LOGATIVO, PATR, LOGQTPLAN, LOGIDADE e LOGIMOVEL). Uma mesma EFPC é representada pelo mesmo formato de ponto

## 4.2 Ajuste do modelo de regressão quantílica longitudinal

Com base nas considerações práticas acerca do tipo de dados que se pretende e na análise exploratória aplicou-se modelo linear misto de regressão quantílica longitudinal, cuja expressão é dada pela equação 10.

As referências na literatura que tratam da modelagem de dados com características semelhantes aos deste trabalho e a análise descritiva preliminar dão sustentação para a transformação logarítmica das variáveis quantitativas e para se considerar efeito aleatório do intercepto, dada a possibilidade de existência de aspectos latentes incorporados no intercepto, como habilidade de gestão ou outros aspectos específicos do individuo observado.

Conforme discutido na Seção 2.2.2 é necessária suposição de distribuição do termo aleatório relativo ao intercepto para estimação dos coeficientes do preditor linear, para fins deste estudo foi considerada a distribuição de Laplace simétrica ($u_i \sim \text{AL}(0, \varphi, \frac{1}{2})$).

Foram estimados os coeficientes do modelo para os quantis 0,1, 0,5 e 0,9 com vistas a se obter uma medida de posicão central da distribuição condicional e avaliar-se o comportamento em pontos mais extremos dessa distribuição.

A definição das variáveis incorporadas no modelo linear teórico baseou-se na minimização do AIC e na avaliação de significância dos parâmetros individualmente que,

consoante Davino Marilena Furno (2014, p. 84), pode ser realizada por meio da estatística T-Student. Nesse sentido, partindo-se de um modelo que contemplava todas as variáveis incluídas na análise (relacionadas na Tabela 1), as variáveis foram sequencialmente eliminadas do modelo, com vistas a minimizar o AIC (método *backward*).

Conforme Karlsson e Hössjer (2022), em modelos mistos a inferência acerca dos valores preditos pode ser realizada marginalmente ou condicionalmente. No primeiro caso, a inferência se dá em nível populacional, calculando-se $\mathbf{x}'_{i,t}\hat{\boldsymbol{\beta}}^{(\tau)}$. Na segunda situação, busca-se uma inferência para determinado agrupamento em que se dá a dependência (o indivíduo), que torna portanto necessário estimar $Q_\tau\left(y_{i,t} \mid u_i, \boldsymbol{\beta}, \mathbf{x}_{it}\right) = u_i + \mathbf{x}'_{i,t}\boldsymbol{\beta}^{(\tau)}$. Por não serem escopo deste trabalho as inferências condicionais, não se abordará em detalhes a estimação de $u_i$, contudo o tema é tratado por Geraci (2014) sob o aspecto teórico e pode ser operacionalmente obtido por meio do pacote *lqmm* pela função *ranef*.

A Tabela 2 apresenta as estimativas pontuais dos parâmetros fixos $\hat{\boldsymbol{\beta}}^{(\tau)}$ referentes às variáveis que foram retidas no modelo final e os respectivos p-valores dos testes de nulidade .

Tabela 2: Coeficientes fixos estimados e p-valor do teste T-Student para a nulidade do parâmetro das variáveis incluídas no modelo final

| | Quantil 0,1 | | Quantil 0,5 | | Quantil 0,9 | |
|---|---|---|---|---|---|---|
| **Parâmetro** | **Coeficiente** | **p-valor** | **Coeficiente** | **p-valor** | **Coeficiente** | **p-valor** |
| Intercepto | 3,824 | <0,0001 | 3,831 | <0,0001 | 3,834 | <0,0001 |
| LOGATIVO | 0,566 | <0,0001 | 0,600 | <0,0001 | 0,602 | <0,0001 |
| PATR | 0,539 | <0,0001 | 0,555 | <0,0001 | 0,554 | <0,0001 |
| LOGQTPLAN | 0,153 | 0,0011 | 0,201 | <0,0001 | 0,193 | <0,0001 |
| LOGIDADE | -0,083 | 0,2641 | -0,207 | 0,0091 | -0,197 | 0,0086 |
| LOGIMOVEL | 0,066 | <0,0001 | 0,052 | 0,0014 | 0,036 | 0,0090 |

A título ilustrativo, descreve-se por meio equação 14 a expressão referente ao preditor linear marginal da mediana da variável LOGDESP condicionada aos valores das cováriaveis ($\hat{q}_{0,5}(LOGDESP)$)

$$\begin{aligned}\hat{q}_{0,5}(LOGDESP) = 3,831 + 0,600LOGATIVO \\ +0,555PATR + 0,201LOGPLAN \\ -0,207LOGIDADE + 0,052LOGIMOVEL\end{aligned} \tag{14}$$

Nota-se por meio análise da Tabela 2 que algumas das observações da análise

exploratória se confirmaram. Primeiramente, observa-se que covariáveis LOGATIVO e PATR de fato apresentam efeitos relevantes. Adicionalmente, verifica-se que ainda que efeito de LOGIDADE não seja significativo ao nível de 5% para o quantil 0,1, para os demais quantis se mostrou uma variável relevante e em todos os casos a estimativa pontual resultou em valor negativo, ou seja, espera-se que EFPC mais jovens tenham gastos administrativos superiores àqueles incorridos por EFPC mais maduras. Também se confirmou a expectativa de redução do efeito da variável LOGPROVBD, em virtude de potencial multicolinearidade com as demais convariáveis (em especial LOGATIVO), a ponto de ser considerada não significativa e, por consequência, excluída do modelo final.

A variabilidade entre indivíduos, medida pela estimativa da variância do coeficiente de intercepto, para os quantis 0,1, 0,5 e 0,9, foi de 0,98, 0,78 e 0,93, respectivamente. Por sua vez, os Coeficientes de Correlação Intra-classe (ICC) indicam forte grau de dependência intra-indivíduo para os três quantis, justificando assim o uso de modelo longitudinal em detrimento da modelagem transversal. O ICC para os quantis de calda (aproximadamente 0,82 para ambos) foram inferiores aos da mediana (0,90), o que pode sugerir que fatores latentes que dêem causa a despesas extremas não se sustentem ao longo do tempo.

Com base nas Figuras 8 e 9 não se identificam padrões flagrantes de agrupamento dos pontos que sugiram heterocedasticidade dos resíduos. Notam-se *outliers* em ambos os gráficos, a despeito da utilização de modelo misto, com intercepto avaliado como variável aleatória, a fim de refletir por exemplo variáveis latentes que impactem o padrão basal de despesas. Na Figura 9 é possível observar a menor uma maior concentração de pontos na região mais central da distribuição condicional dos resíduos, que pode ser contudo efeito da menor densidade de pontos nessa região.

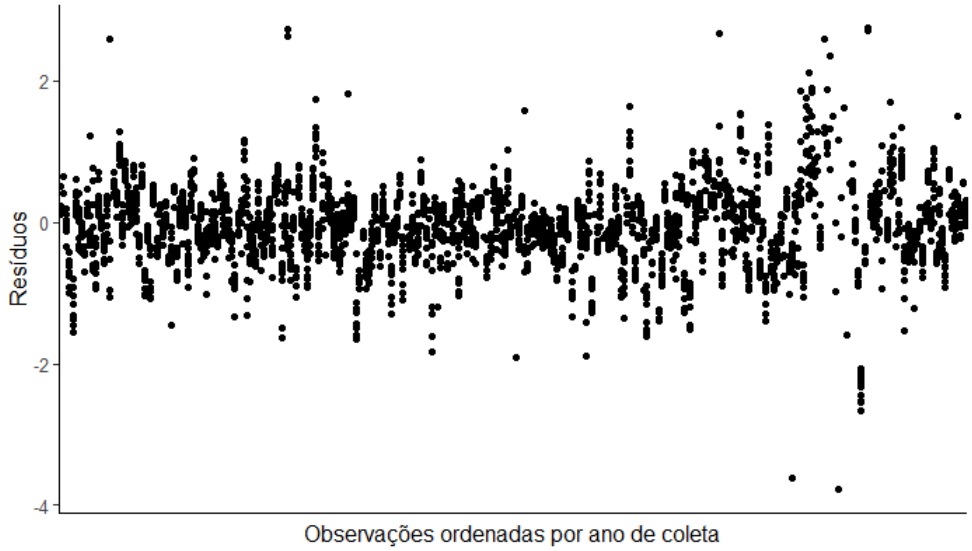


Figura 8: Gráfico de dispersão dos resíduos em relação às observações ordenadas por tempo

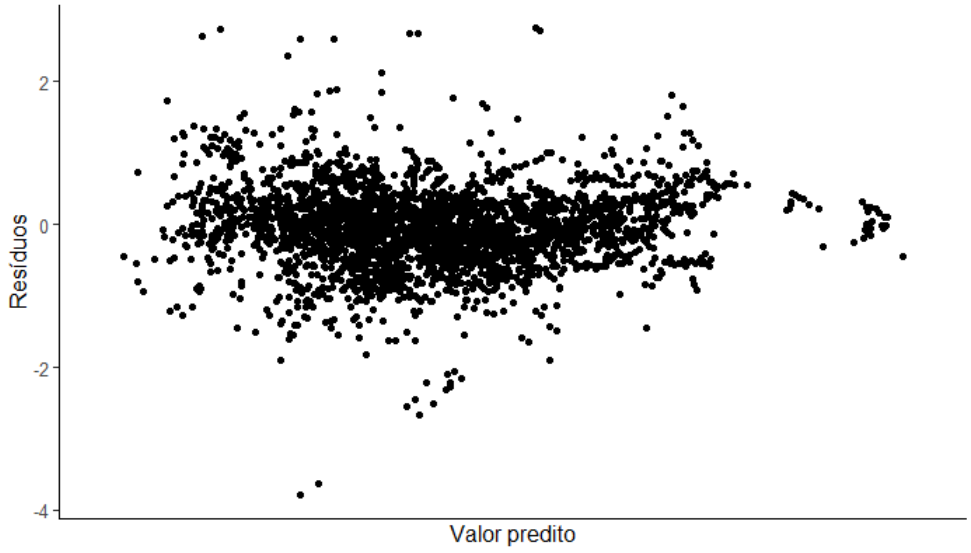


Figura 9: Gráfico de dispersão dos resíduos em relação aos valores preditos pelo modelo final

Por fim, avaliou-se a possibilidade de que o efeito da variável LOGATIVO, que se mostrou mais significativa no modelo, fosse aleatório, ou seja, de que aspectos individuais levassem a diferentes níveis de sensibilidade a LOGDESP dessa covariável. Aplicou-se, dessa maneira, generalização do modelo de regressão quantílica misto desenvolvido em (GERACI; BOTTAI, 2014), que prevê a aplicação de efeitos aleatórios não só para o intercepto, como também para os coeficientes angulares. O AIC desse modelo foi aproximadamente o dobro daqueles testados no processo de seleção do modelo final, que consideravam somente aleatoriedade do intercepto. Dessa forma não foram realizadas maiores

investigações acerca de possível aleatoriedade de outros efeitos.

# 5 Conclusão

A aplicação de modelo de regressão quantílica linear misto proposto por Geraci e Bottai (2007) para modelar as despesas administrativas de fundos de pensão se mostrou adequada. A presença relativamente alta de valores extremos (*outliers*) favorece o uso de métodos de estimação de parâmetros que sejam menos sensíveis a essas ocorrências, tal qual a função de perda da regressão quantílica. A transformação logarítmica das variáveis quantitativas do modelo, como forma de linearizar a estrutura funcional demonstrou fundamentos do ponto de vista teórico e prático, o que pode ser conveniente para a interpretação do modelo, dado que modelos aditivos facilitam a compreensão dos efeitos de forma isolada.

Avalia-se que existe potencial no uso do modelo para fins de avaliação do grau de ineficiência na gestão de gastos dos fundos de pensão. É possível analisar as fontes de variabilidade do modelo, o termo de erro e o intercepto aleatório, as quais podem servir como medida para essa avaliação, caso se suponha que sejam resultantes predominantemente de aspectos latentes que possam ser associados a maior ou menor qualidade na gestão administrativa.

Ainda nesse sentido, os próprios efeitos das variáveis explicativas do modelo podem ser investigados. Identificadas covariáveis cujos efeitos não possam ser "justificados" sob o aspecto econômico-financeiro, é possível comparar os valores dos quantis condicionais preditos a partir dos dados observados com a predição que leve em conta ajuste nos valores dessas variáveis. Em particular, chama-se a atenção para o efeito da variável explicativa dicotômica incluída no modelo final referente ao tipo patrocínio de patrocínio predominante da EFPC (PATR). Nota-se substancial deslocamento da distribuição condicional da despesa para a direita (elevação do valor predito) em face de o patrocínio da EFPC ser realizado por ente de natureza pública.

Embora tenham sido feitas avaliações incipientes relativamente à aplicabilidade de modelos alternativos, há espaço para aprofundar essas investigações. Também se destaca que a inclusão de novas covariáveis, disponíveis em bancos de dados ou mesmo que demandem coleta localmente, podem vir a melhorar a qualidade do modelo.

# Referências

BOZDOGAN, H. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, Springer, v. 52, n. 3, p. 345–370, 1987.

BRASIL. Lei complementar nº 109, de 29 de maio de 2001. *Diário Oficial [da] República Federativa do Brasil*, 2001.

BREUSCH, T. S.; PAGAN, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, JSTOR, p. 1287–1294, 1979.

DAVINO MARILENA FURNO, D. V. C. *Quantile regression : theory and applications*. [S.l.: s.n.], 2014.

DICK, A.; POMORSKI, L. Is bigger better? size and performance in pension plan management. *Rotman School of Management Working Paper*, 2010.

GERACI, M. Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, v. 57, n. 13, p. 1–29, 2014.

GERACI, M.; BOTTAI, M. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, Oxford University Press, v. 8, n. 1, p. 140–154, 2007.

GERACI, M.; BOTTAI, M. Linear quantile mixed models. *Statistics and Computing*, v. 24, n. 3, p. 461–479, 2014.

GOLDFELD, S. M.; QUANDT, R. E. The estimation of cobb-douglas type functions with multiplicative and additive errors. *International Economic Review*, JSTOR, v. 11, n. 2, p. 251–257, 1970.

HSIAO, C. *Analysis of panel data*. [S.l.]: Cambridge university press, 1986.

HUANG HANZE ZHANG, J. C. Q.; HE, M. Quantile regression models and their applications: A review. *Piracicaba: USP*, 2011.

KARLSSON, M.; HÖSSJER, O. A comparison between quantile regression and linear regression on empirical quantiles for phenological analysis in migratory response to climate change. *arXiv preprint arXiv:2202.02206*, 2022.

KOENKER, R.; JR, G. B. Regression quantiles. *Econometrica: journal of the Econometric Society*, JSTOR, p. 33–50, 1978.

MARINO, M. F.; FARCOMENI, A. Linear quantile regression models for longitudinal experiments: an overview. *Metron*, Springer, v. 73, n. 2, p. 229–247, 2015.

MARX, B. D.; EILERS, P. H. Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, Elsevier, v. 28, n. 2, p. 193–209, 1998.

PREVIC. Relatório das despesas administrativas das efpc. *https://www.gov.br/economia/pt-br/orgaos/entidades-vinculadas/autarquias/previc/centrais-de-conteudo/noticias/previc-divulga-relatorio-sobre-as-despesas-administrativas-da-efpcs*, 2021.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: ⟨https://www.R-project.org/⟩.

SCHMIDT, D.; GERMANO, A. M.; MILANI, T. L. Subjective sensitivity data: Considerations to treat heteroscedasticity. *Cogent Medicine*, Cogent OA, v. 6, n. 1, p. 1673086, 2019. Disponível em: ⟨https://doi.org/10.1080/2331205X.2019.1673086⟩.