
Universidade Federal do Rio de Janeiro

# Modelo de previsão para a premissa de variação dos custos médicos e hospitalares (VCMH), utilizada na mensuração do passivo com planos de saúde empresariais.

André Monteiro Dovalski

2015

UFRJ

# Modelo de previsão para a premissa de variação dos custos médicos e hospitalares (VCMH), utilizada na mensuração do passivo com planos de saúde empresariais.

André Monteiro Dovalski

Monografia apresentada ao Instituto de Matemática e ao Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para a conclusão dos cursos de Ciências Atuariais e Estatística bem como obtenção dos respectivos títulos de Bacharelado.

Orientador: Ralph dos Santos Silva

Rio de Janeiro, 20 de outubro de 2015.

# Modelo de previsão para a premissa de variação dos custos médicos e hospitalares (VCMH), utilizada na mensuração do passivo com planos de saúde empresariais.

André Monteiro Dovalski

Orientador: Ralph dos Santos Silva

Monografia apresentada ao Instituto de Matemática e ao Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para a conclusão dos cursos de Ciências Atuariais e Estatística bem como obtenção dos respectivos títulos de Bacharelado.

_____

Prof. Ralph dos Santos Silva

IM-UFRJ

_____

Prof. Thaís Cristina de Oliveira da Fonseca

IM-UFRJ

_____

Sandro de Azambuja

Rio de Janeiro, 20 de outubro de 2015.

Dovalski, André Monteiro

Modelo de previsão para a premissa de variação dos custos médicos e hospitalares, utilizada na mensuração do passivo com planos de saúde empresariais / André Monteiro Dovalski – Rio de Janeiro: UFRJ/IM, 2015.

iii, 26 f.: il.; 31 cm.

Orientador: Ralph dos Santos Silva

Monografia – UFRJ/ IM/ Graduação em Ciências Atuariais e Estatística, 2015.

Referências Bibliográficas: f. 27

1. VCMH. 2. Custos Médicos. 3. Planos de saúde. 4. Regressão Linear I. Silva, Ralph dos Santos. II. Universidade Federal do Rio de Janeiro, Instituto de Matemática. III. Modelo de previsão para a premissa de variação dos custos médicos e hospitalares (VCMH), utilizada na mensuração do passivo com planos de saúde empresariais.

**RESUMO**

Modelo de previsão para a premissa de variação dos custos médicos e hospitalares (VCMH), utilizada na mensuração do passivo com planos de saúde empresariais.

André Monteiro Dovalski

Orientador: Ralph dos Santos Silva

Monografia apresentada ao Instituto de Matemática e ao Departamento de Métodos Estatísticos da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para a conclusão dos cursos de Ciências Atuariais e Estatística bem como obtenção dos respectivos títulos de Bacharelado.

O trabalho apresentado tem como objetivo realizar a previsão para o próximo ano da variação dos custos médicos e hospitalares (VCMH), uma premissa utilizada no cálculo das provisões matemáticas com benefícios pós-emprego de assistência médica. Utilizamos um modelo de mistura para a modelagem devido à alta quantidade de zeros na população. Para a parte nula do modelo, estimamos a proporção de zeros. Já para a parte estritamente positiva do modelo realizamos dois modelos de regressão: um primeiro mais complexo utilizando diversas covariáveis e um segundo mais simples utilizando apenas uma função temporal. Realizamos diversos procedimentos visando analisar os resultados encontrados e sua aplicabilidade.

*Palavras-chave*: Regressão linear, benefício pós-emprego, assistência médica, custo médicos, variação dos custos médicos e hospitalares, premissa, modelo de mistura.

**Para meus pais**

**Cristina Maria Sousa Monteiro**

**E**

**José Veriano Dovalski**

**Para minhas avós**

**Maria Alice Sousa Monteiro**

**E**

**Arlete Ramos Dovalski**

## AGRADECIMENTOS

Ao professor e orientador Ralph dos Santos Silva, por todo o tempo e ajuda dedicados a este projeto.

Ao colega de profissão Sandro de Azambuja pela ideia em começar esse projeto e este estudo tão enriquecedores.

Aos amigos que estiveram comigo em toda a trajetória dentro desta universidade.

## Sumário

## Índice de Tabelas

## Índice de Gráficos

**Capítulo1: Introdução**

O principal objetivo deste trabalho é realizar a previsão da premissa de Variação dos Custos Médicos e Hospitalares, utilizando dados de utilização de um plano de saúde. Para isto iremos introduzir algumas informações e conceitos anteriormente.

Na década de 1960 foram iniciadas extraoficialmente as atividades do setor de saúde suplementar no Brasil. Algumas empresas dos setores industrial e de serviços começaram a oferecer assistência médica a seus empregados, mesmo sem nenhum tipo de regulamentação pública.

O Programa Estadual de Defesa do Consumidor, originado pelo Código de Defesa do Consumidor no ano de 1990, começou a receber diversos tipos de reclamações relativos às empresas privadas de saúde. Este fato levou a necessidade da criação de um órgão regulador para esse segmento da economia. No ano de 2000, foi criada pela Lei nº9961, a Agência Nacional de Saúde Suplementar. Seu objetivo é realizar a regulamentação de uma atividade muito complexa, utilizando como principal ferramenta, a Lei nº9656/1998 que fora sancionada dois anos antes pelo Congresso nacional.

A Lei nº9656[1] é um marco na legislação brasileira, pois não somente dispõe de instruções regulamentares para os planos privados, como também para a assistência à saúde, oferecida a empregados muitas vezes pós-aposentadoria.

No caso da assistência a saúde oferecida pelas empresas a seus empregados, os artigos 30 e 31 da Lei nº9656 são de suma importância para o que iremos discutir no corpo deste trabalho. Com eles e com a aprovação pela Comissão de Valores Mobiliários (CVM) do Comitê de Pronunciamentos Técnicos 33 (CPC 33) em 2009, houve a necessidade de calcular uma provisão matemática para os participantes que fossem elegíveis. O CPC 33 dispõe de parâmetros técnicos para a contabilização e divulgação dos benefícios oferecidos a empregados. Basicamente o artigo 30, infere que todo o empregado que contribuir efetivamente com o plano de assistência médica oferecido pela empresa, estará habilitado a permanecer no plano quando de seu desligamento por pelo menos 6 meses e no máximo 2 anos, desde que pague também a parcela patronal, seguindo algumas condições. Já o artigo

[1] http://www.planalto.gov.br/ccivil_03/Leis/L9656.htm

31, decreta que todo aposentado que contribuir com plano de saúde decorrente de vínculo empregatício por pelo menos 10 anos, tem direito a levar este benefício para o resto da vida com as mesmas condições desde que arque com o custo total da contribuição e também dispõe mais alguns detalhes.

A provisão matemática calculada para este tipo de benefício pós-emprego leva em conta diversas hipóteses atuariais, como por exemplo: tábuas de mortalidade e/ou rotatividade. A premissa de Variação dos Custos Médicos e Hospitalares (VCMH) é fundamental na estimação do passivo com o benefício pós-emprego de assistência médica, pois ela é utilizada na projeção dos fluxos de gasto com cada participante presente no plano de benefícios.

### 1.1. Objetivo

O corpo deste trabalho traz como objetivo a apresentação de uma metodologia de previsão para a mensuração da premissa de Variação dos Custos Médicos e Hospitalares de uma determinada companhia. Para isso utilizaremos um modelo de mistura, visto que os dados utilizados tem um grande número de zeros, este modelo será explicado mais a frente.

O restante do presente trabalho é composto por três principais capítulos. No capítulo 2, metodologia, apresentaremos de forma resumida as bases técnicas estatísticas utilizadas. No capítulo 3, aplicação, descreveremos o modelo apresentando os dados utilizados e finalizaremos com a descrição dos resultados obtidos. Por fim, na conclusão, capítulo 4, teremos as considerações finais e demais comentários.

## Capítulo 2: Metodologia

Neste capítulo iremos apresentar, resumidamente, todas as bases técnicas estatísticas utilizadas para a realização do objeto de estudo presente neste trabalho.

### 2.1. Modelo de Mistura

Iremos considerar a função de probabilidade dada por $g(y) = I(y = 0)$

E também a função de probabilidade $h(y)$ dada por:

$$h(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\ln(y) - \mu)^2\right\} I(y > 0).$$

Vamos definir também uma variável binária $Z \sim Bern(\alpha)$, portanto $P(Z = 1) = \alpha$ e $P(Z = 0) = 1 - \alpha$, onde $Z = 1$ significa que o valor encontrado na variável y é zero.

Portanto, iremos escrever nosso modelo de mistura como:

$$f(y|\mu, \sigma^2, \alpha) = [\alpha g(y) + (1 - \alpha)h(y|\mu, \sigma^2)]\, I(y \geq 0)\ com\ 0 \leq \alpha \geq 1,$$

onde

- $g(y) = f(y|Z = 1, \mu, \sigma^2)$ e;
- $h(y|\mu, \sigma^2) = f(y|Z = 0, \mu, \sigma^2)$.

É interessante utilizarmos este tipo de modelo quando temos dentro dos dados analisados, dois tipos de características bem distintas. No caso aqui apresentado, iremos utilizá-lo em virtude de os dados apresentarem uma grande quantidade de zeros. Portanto será modelada a parte nula de nossa variável de interesse, separadamente da parte estritamente positiva.

### 2.2. Regressão Linear

O modelo de regressão linear é utilizado para estudarmos a relação entre uma variável dependente com uma ou mais variáveis independentes. A forma genérica da regressão linear é definida como:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \varepsilon,$$

sendo,

- $y$ a variável dependente ou a ser explicada;
- $x_1 \dots x_k$ as variáveis independentes ou regressoras;
- $\beta_0$ o intercepto;
- $\beta_1 \dots \beta_k$ os coeficientes das variáveis regressoras e;
- $\varepsilon$ o erro aleatório com média 0 (zero) e variância constante (homocedasticidade).

As variáveis regressoras ou independentes podem se de qualquer tipo, seja ele quadrático, binário, com interações ou sem, etc. O fato que indica a linearidade do modelo está em como os betas se relacionam com as covariáveis e a variável a ser explicada. Utilizaremos esta técnica para modelar a parte positiva dos dados.

Em geral, assumimos a hipótese de normalidade do erro aleatório. No entanto, para ajustes com estimação pelo método dos mínimos quadrados não existe a necessidade de se impor uma distribuição para o erro. Só precisamos que sejam satisfeitas duas condições: média zero e variância constante. Rotineiramente utilizamos a premissa da normalidade para o erro, para a realização de testes de hipótese ou quando desejamos utilizar o método de estimação por máxima verossimilhança. Podemos também, deixar o ajuste do modelo mais robusto, ou seja, que podemos inserir outras distribuições como, por exemplo, uma que tenha caudas mais pesadas que a Normal, como a t-Student. Há também a possibilidade de utilizarmos uma outra abordagem (que não foi aplicada neste trabalho) onde permitimos que o erro aleatório seja oriundo de uma distribuição assimétrica. Para mais detalhes sobre modelos de regressão veja GREENE (2012).

### 2.2.1. Previsão para um modelo de regressão linear

Para realizarmos a previsão de um valor $y^0$ associado a um vetor de variáveis regressoras $x^0$ precisamos construir este vetor e efetuar a regressão novamente. O valor previsto será:

$$y^0 = x^{0\prime}\beta + \varepsilon^0, \text{ sendo } \varepsilon \sim N(0, \sigma^2).$$

Os $\beta$'s da regressão linear e de sua previsão podem ser estimados por vários métodos, e apresentaremos dois principais ainda neste capítulo.

### 2.2.2. Intervalos de Predição

Após a estimação do valor da previsão seu intervalo de confiança será dado por:

$$IC_{100(1=\alpha)\%}(y^0) = (\hat{y}^0 - t_{\frac{\alpha}{2};n-p} \,.\, ep(e^0), \hat{y}^0 + t_{\frac{\alpha}{2};n-p} \,.\, ep(e^0)) \,,$$

sendo

- $y^0$ o valor previsto;
- $ep(e^0)$ é o erro padrão da previsão;
- $\alpha$ o nível de confiança do intervalo; e
- $t_{\frac{\alpha}{2};n-p}$ o quantil correspondente da distribuição t-Student com $n-p$ graus de liberdade.

### 2.2.3. Estimação por Mínimos Quadrados

Seja a representação vetorial a seguir do modelo de regressão:

$$y_i = x_i'\beta + \varepsilon_i.$$

Para utilizarmos a estimação por mínimos quadrados precisamos minimizar o vetor da soma de quadrados dos resíduos:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - x_i' b_0)^2,$$

Após isso encontraremos o estimador de mínimos quadrados para $\beta = b_0 = (X'X)^{-1}X'y$. (GREENE, 2012)

### 2.2.4. Estimação por Máxima Verossimilhança

Primeiramente precisamos definir a função de máxima verossimilhança (GREENE, 2012).

Seja $y = (y_1, \dots, y_n)'$ uma amostra aleatória independente e identicamente distribuída e $\theta$ um conjunto de parâmetros. A função de máxima verossimilhança é dada por:

$$f(y_1, \dots, y_n \mid \theta) = \prod_{i=1}^{n} f(y|\theta) = L(\theta|y).$$

Assim, os estimadores de máxima verossimilhança são encontrados pela maximização da função de verossimilhança descrita acima. Portanto:

$$\hat{\theta}_{EMV} = \max_{\theta} L(\theta|y).$$

Em geral, maximizamos o logaritmo da função de máxima verossimilhança definida por:

$$\ln\big(L(\theta|y)\big) = l(\theta|y)\ .$$

Para estimar as variâncias e os erros padrões dos estimadores de máxima verossimilhança, utilizamos o inverso da matriz de informação de Fisher $I(\theta) = (-E_0[\frac{\partial^2 lnL}{\partial\theta_0\partial\partial_0'}])^{-1}$. Quando o cálculo anterior não for possível de ser realizado, utilizamos o negativo do inverso da derivada segunda do logaritmo da função de máxima verossimilhança. Nesse caso precisamos utilizar o Algoritmo abaixo para realizar a estimação por máxima verossimilhança.

### 2.3. Algoritmo Esperança-Maximização

O algoritmo esperança-maximização é utilizado para estimarmos - por máxima verossimilhança - os parâmetros de um modelo quando temos dados incompletos ou quando a solução analítica para a maximização desta é inviável. Neste caso, em virtude da quantidade de zeros nos dados e também da magnitude dos valores, optamos por utilizá-lo. Quando o aplicamos, este se reduziu a um caso particular, onde precisávamos apenas calcular a proporção de zeros em nossos dados.

Para isso, precisamos calcular os dois passos das iterações do algoritmo. O passo E, onde se calcula a esperança e o M onde esta é maximizada. Logo, de forma geral avaliamos o valor esperado (passo E), $E[l^c(\theta|y)|\theta^0]$, e depois encontramos $\theta^1$ (passo M) que maximiza este valor esperado.

Seja $Z$ uma variável aleatória, onde $Z{\sim}Bern(\alpha)$, temos $P(Z=1) = \alpha\ e\ P(Z=0) = 1-\alpha$ tal que $E(Z) = \alpha$.

Assim podemos escrever o modelo da seguinte forma:

$f(y|Z=1,\theta) = g(y)$ e $f(y|Z=0,\theta) = h(y|\theta)$.

Portanto, para calcular o máximo da função de verossimilhança precisamos calcular:

$E(l^c(\theta,\alpha|y,z)|y,\theta_0,\alpha_0)$ e para isto é necessário avaliar $\pi = P(Z=1|y,\theta_0,\alpha_0)$.

Logo, como $\pi = \begin{cases} 1 & y=0 \\ 0 & y>0 \end{cases}$, vamos definir $n_1 = \sum_{i=1}^{n} I(y_i = 0)$ e $n_2 = \sum_{i=1}^{n} I(y_i > 0)$, consequentemente $n_1 + n_2 = n$ e ainda que $y_j^*$ seja o vetor da amostra onde $y > 0$ tal que $(y_1^*, \dots y_{n_2}^*)$. Dito isto, a maximização de $\pi$ se dá da seguinte forma: $\hat{\pi} = \frac{n_1}{n_1+n_2}$.

Logo, teremos que para o passo E, será necessário calcularmos:

$$\pi = \begin{cases} 1 & y=0 \\ 0 & y>0 \end{cases} \text{ e}$$

$$E(l^c(\theta,\alpha|y,z)|y,\theta_0,\alpha_0) = n_1 \ln(\alpha) + n_2 \ln(1-\alpha) + \sum_{i=1}^{n} \ln(h(y_j^*|\theta).$$

E para o passo M;

$$\hat{\pi} = \frac{n_1}{n_1+n_2} \text{ (não se altera) e}$$

$$\hat{\beta} = (X'X)^{-1}X'y^* \text{ e } \hat{\sigma}^2 = \frac{e\prime e}{n_2} \text{ com } e = y^* - X\hat{\beta}.$$

### 2.4. Testes para Normalidade

#### 2.4.1. Testes de Assimetria e Curtose

Utilizamos dois testes t-Student para avaliarmos se há normalidade nos dados.

Para assimetria amostral, testamos:

$$t = \frac{A}{\sqrt{\sigma/n}} \approx N(0,1),$$

sendo

- A o coeficiente de assimetria da amostra;
- n o número de observações e;
- $\sigma$ o desvio padrão amostral.

Rejeitaremos a hipótese de normalidade caso a estatística |t| seja maior que o percentil 100(1-α/2) da normal padrão, sendo α o nível de significância escolhido.

Para testarmos o excesso de curtose, utilizamos:

$$t = \frac{C-3}{\sqrt{24/n}} \approx N(0,1),$$

sendo C o coeficiente de curtose da amostra e n o número de observações.

Também rejeitaremos a hipótese nula (normalidade) caso a estatística |t| seja maior que o percentil 100(1- α/2) da normal padrão, sendo α o nível de significância escolhido.

### 2.4.2. Teste de Jarque-Bera

O teste de Jarque-Bera utiliza a estatística J, para testar se a distribuição segue normalidade. Este é baseado em uma combinação entre os coeficientes de curtose e assimetria amostrais. Calculamos a estatística da seguinte forma:

$$J = n\left[\frac{A^2}{6} + \frac{(C-3)^2}{24}\right],$$

sendo n o número de observações, C o coeficiente de curtose da amostra e A o coeficiente de assimetria da amostra.

Para avaliarmos se a hipótese nula é satisfeita, o excesso de curtose (C-3) e de assimetria (A) devem ser próximos de zero, pois sob normalidade o coeficiente de curtose é igual a 3 e o de assimetria, igual a zero.

### 2.4.3. Teste de Shapiro-Wilk

O teste de Shapiro-Wilk, proposto em 1965, é baseado no cálculo de uma estatística $W$ para confrontar a hipóteses nula $H_0$, onde os dados da amostra seguem uma distribuição normal e a hipótese alternativa $H_1$, onde estes dados não seguem uma distribuição normal. Rejeitaremos a hipótese nula para valores pequenos de $W$.

A expressão abaixo mostra o cálculo da estatística $W$:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_{(i)}-\bar{x})^2},$$

sendo

- $x_{(i)}$ os valores da amostra ordenados; e
- $b = \sum_{i=1}^{\frac{n}{2}} a_{n-i+1}\,(x_{(n-i+1)} - x_{(i)})$, para n par ou $b = \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1}\,(x_{(n-i+1)} - x_{(i)})$ para n ímpar.

**2.5. Critério de Informação de Akaike (AIC)**

Para compararmos dois ou mais modelos, utilizamos o Critério de Informação de Akaike (AIC) e escolhemos o modelo com menor valor nessa estatística.

Definimos o AIC como:

$$AIC = -2\ln\left(L(\hat{\theta})\right) + 2p,$$

sendo $p$ o números de parâmetros considerado no modelo e $L$ a função de verossimilhança do modelo avaliada no ponto $\hat{\theta}$ de máximo. Escolhemos como melhor modelo àquele que apresenta menor AIC, isso implica que estaremos escolhendo o modelo com menor soma do quadrado dos resíduos e o menor número de parâmetros (WEISBERG, 2005), o que é desejável.

## Capítulo 3: Aplicação

O cálculo do passivo atuarial referente ao benefício de assistência médica, oferecido por empresas a seus empregados no Brasil e no mundo, tem como uma das principais premissas, a variação dos custos médicos e hospitalares (VCMH). Esta é utilizada na projeção do custo futuro com despesas médicas.

Foi utilizada uma base com dados de 2008 até 2013 e cerca de 215 mil registros presentes em todos os anos. Foi necessário realizar uma comparação entre os registros ano a ano para termos certeza que teríamos a evolução destes no decorrer dos anos. Cada um dos registros representa um participante de um plano de assistência médica oferecido por uma empresa cujo nome deve ser mantido em sigilo. Para cada um dos associados, a base de dados apresenta o valor das despesas em procedimentos médicos (custeados pela empresa) divididos em três principais categorias: Pequeno Risco, que compreende toda a gama de exames e consultas médicas, além de pequenos procedimentos. Grande Risco, que engloba as internações e operações hospitalares e por fim, o grupo Odontologia, onde ficam representados todos os gastos com tratamentos dentais em geral. Além disso, os dados foram mascarados pela multiplicação de uma constante para que os reais não sejam divulgados.

A variável modelada foi a Despesa Total, que foi obtida com a soma dos valores das despesas de Pequeno Risco, Grande Risco e Odontologia para cada participante. Com ela nós conseguiremos realizar a previsão do valor do total de despesa para o ano seguinte e assim obter a VCMH.

Após realizarmos uma análise exploratória da base de dados, verificamos que muitos participantes não utilizaram a assistência médica em determinadas categorias e por ano. Levando assim a uma quantidade muito elevada de zeros nos dados e consequentemente uma assimetria da distribuição dos valores dos gastos como veremos a seguir:

### 3.1. Análise Exploratória dos Dados

Tabela 1: Descrição das Variáveis.

| Variável | Média | Mediana | Assimetria | Curtose | Desvio Padrão |
|---|---|---|---|---|---|
| Idade | 42,41 | 45 | -0,05585 | -0,9408 | 21,87 |
| Salário | 597,8 | 517,6 | 1,617 | 4,126 | 377,6 |
| Despesa Total | 147,2 | 18,19 | 25,96 | 1290 | 918,6 |

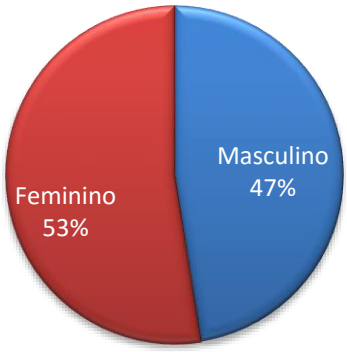


Gráfico 1: Proporção entre Gêneros.


Gráfico 2: Proporção de valores nulos.

Apesar dos 14% aparentarem um valor relativamente baixo, este representa 29.534 registros da amostra. Isso levou-nos a um problema de modelagem principalmente por questões computacionais, pois temos uma quantidade significativa de valores nulos.

Para tratar esta questão, optamos por utilizar um modelo de mistura, como descrito no capítulo 2: Metodologia.

Utilizamos a estimação por máxima verossimilhança através do algoritmo EM. Ao final de sua aplicação, este se reduziu a calcularmos a proporção de zeros na amostra como já havíamos realizado e a considerar um modelo de regressão somente para os valores estritamente positivos dos dados.

### 3.2. Construção do Modelo e Apresentação de Resultados

Com respeito a parte com observações estritamente positiva do modelo, começamos realizando um ajuste por meio de uma regressão linear com erros normais, além de realizarmos a transformação logarítmica na variável dependente devido ao seu grande valor numérico. Essa transformação também ajudou a melhorar a simetria dos dados e a variabilidade do erro se aproximando mais da hipótese de homocedasticidade. Utilizamos o método *backward* para definir qual seria o melhor modelo a ser utilizado. Este método consiste em, primeiramente, realizar o ajuste da regressão com todas as variáveis e interações possíveis. Após isso, a medida que as variáveis ou interações se mostrarem não significativas, iremos retirá-las do modelo uma por uma, sempre refazendo o ajuste afim de verificar sua contribuição para explicação da variável de interesse. Por fim, chegamos ao que será considerado o melhor modelo, apenas com as variáveis e interações que se mostrarem estatisticamente significativas com nível de pelo menos 5%.

Tabela 2: Descrição da Variável Log Despesa Total Positiva.

| Variável | Média | Mediana | Assimetria | Curtose | Desvio Padrão |
|---|---|---|---|---|---|
| Log Despesa Total Positiva | 3,24 | 3,19 | 0,138 | 0,8292 | 1,78 |

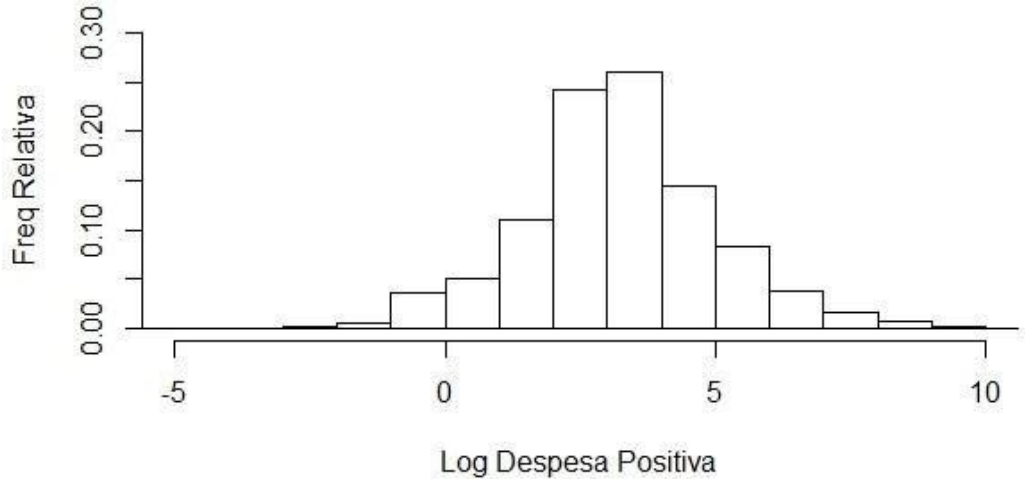


Gráfico 3: Histograma do Log da Despesa Total Positiva.

Vemos que o valor do coeficiente de assimetria é positivo, ou seja, a distribuição dos dados é mais concentrada à direita, o que está representado graficamente pelo histograma. Apesar disto, seguimos com o modelo de regressão linear normal.

Para o nosso primeiro ajuste, considerando todas as variáveis e suas possíveis interações, obtivemos os seguintes resultados:

Tabela 3: Coeficientes estimados – 1º ajuste Normal. Os dois pontos indicam interações entre as variáveis.

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | 2,65000 | 0,12200 | 21,74 | $2x10^{-16}$ |
| Sexo | -0,16200 | 0,03250 | -4,98 | $6,5x10^{-7}$ |
| Idade | 0,01820 | 0,00042 | 43,17 | $2x10^{-16}$ |
| Salário | 0,00002 | 0,00003 | 0,85 | 0,3967 |
| Tempo | -1,30000 | 0,25000 | -5,22 | $1,8x10^{-7}$ |
| $Tempo^2$ | 1,07000 | 0,18200 | 5,89 | $3,9x10^{-9}$ |
| $Tempo^3$ | -0,37000 | 0,06020 | -6,15 | $7,8x10^{-10}$ |
| $Tempo^4$ | 0,05860 | 0,00916 | 6,39 | $1,6x10^{-10}$ |
| $Tempo^5$ | -0,00342 | 0,00052 | -6,53 | $6,4x10^{-11}$ |
| Sexo:Idade | 0,00331 | 0,00063 | 5,22 | $1,8x10^{-7}$ |
| Sexo:Salário | 0,00001 | 0,00004 | 0,25 | 0,7996 |
| Idade:Salário | $1x10^{-6}$ | $1x10^{-6}$ | 2,17 | 0,0299 |
| Sexo:Tempo | -0,03600 | 0,00850 | -4,24 | 0,00002 |
| Idade:Tempo | 0,00013 | 0,00011 | 1,2 | 0,2283 |
| Salário:Tempo | -0,00003 | 0,00001 | -4,27 | 0,00002 |
| Sexo:Idade:Salário | $1x10^{-6}$ | $1x10^{-6}$ | -2,06 | 0,0393 |
| Sexo:Idade:Tempo | 0,00076 | 0,00016 | 4,69 | $2,7x10^{-6}$ |
| Sexo:Salário:Tempo | 0,00003 | 0,00001 | 3,02 | 0,0025 |
| Idade:Salário:Tempo | $1x10^{-6}$ | $1x10^{-6}$ | 0,07 | 0,943 |
| Sexo:Idade:Salário:Tempo | $1x10^{-6}$ | $1x10^{-6}$ | -2,35 | 0,019 |

Ao analisarmos os resultados, podemos concluir que a variável salário não se mostrou estatisticamente significativa ao nível de 5%.

O valor das estatísticas $R^2$ e $R^2$ ajustado foi de 9,3%, ou seja, as variáveis explicativas ajudaram a entender somente este percentual da variabilidade da variável independente.

Realizamos também o teste F, para confrontarmos as hipóteses de que os coeficientes estimados são todos iguais a zero (hipótese nula) contra pelo menos um deles é estatisticamente diferente de zero (hipótese alternativa). O P-valor encontrado para o teste foi muito próximo de zero, indicando a rejeição da hipótese nula.

Dessa forma, para pouparmos memória e simplificarmos o modelo que estamos construindo, decidimos por eliminar a variável salário e suas interações e prosseguir com a modelagem.

Os resultados do segundo ajuste são apresentados na tabela a seguir:

Tabela 4: Coeficientes estimados - 2º ajuste Normal. Os dois pontos indicam interações entre as variáveis.

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | 2,68000 | 0,12100 | 22,24 | $2x10^{-16}$ |
| Sexo | -0,17600 | 0,01650 | -10,67 | $2x10^{-16}$ |
| Idade | 0,01850 | 0,00023 | 81,12 | $2x10^{-16}$ |
| Tempo | 1,33000 | 0,25000 | -5,31 | $1,1x10^{-7}$ |
| $Tempo^2$ | 1,07000 | 0,18200 | 5,89 | $3,9x10^{-9}$ |
| $Tempo^3$ | -0,37000 | 0,06020 | -6,15 | $7,7x10^{-10}$ |
| $Tempo^4$ | 0,05860 | 0,00916 | 6,39 | $1,6x10^{-10}$ |
| $Tempo^5$ | -0,00342 | 0,00052 | -6,53 | $6,4x10^{-11}$ |
| Sexo:Idade | 0,00279 | 0,00034 | 8,23 | $2x10^{-16}$ |
| Sexo:Tempo | -0,00892 | 0,00434 | -2,06 | 0,0398 |
| Idade:Tempo | 0,00031 | 0,00006 | 5,29 | $1,2x10^{-7}$ |
| Sexo:Idade:Tempo | 0,00025 | 0,00009 | 2,89 | 0,0038 |

Podemos identificar que não houve variáveis que não fossem estatisticamente significativas ao nível de 5%. Apesar disso, como esperado para este modelo, foi constatada uma leve redução no valor das estatísticas $R^2$ e $R^2$ ajustado que foi para 9,26%.

Também realizamos aqui o teste F, onde o P-valor encontrado foi próximo de zero indicando a rejeição da hipótese nula.

Desta forma, não há motivo para retirarmos mais nenhuma variável ou interação do modelo. Prosseguiremos com as análises dos resíduos da regressão.

Tabela 5: Testes para normalidade dos resíduos.

| Testes para Normalidade dos Resíduos | P-valor |
|---|---|
| T (Assimetria) | 0 |
| T (Excesso de Curtose) | 0 |
| Jarque-Bera | $2x10^{-16}$ |
| Shapiro Wilk | $2x10^{-15}$ |

Pelos testes realizados, inferimos pelo p-valor que não há presença de normalidade dos resíduos. Para tornar mais fácil a visualização da não normalidade, apresentamos os gráficos dos resíduos a seguir:

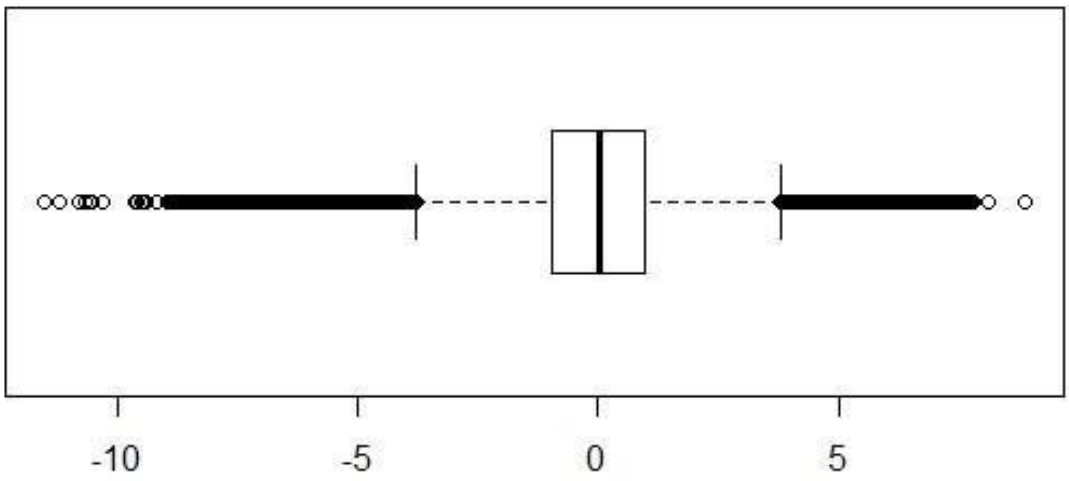


Gráfico 4: Boxplot dos Resíduos – 2º Ajuste Normal

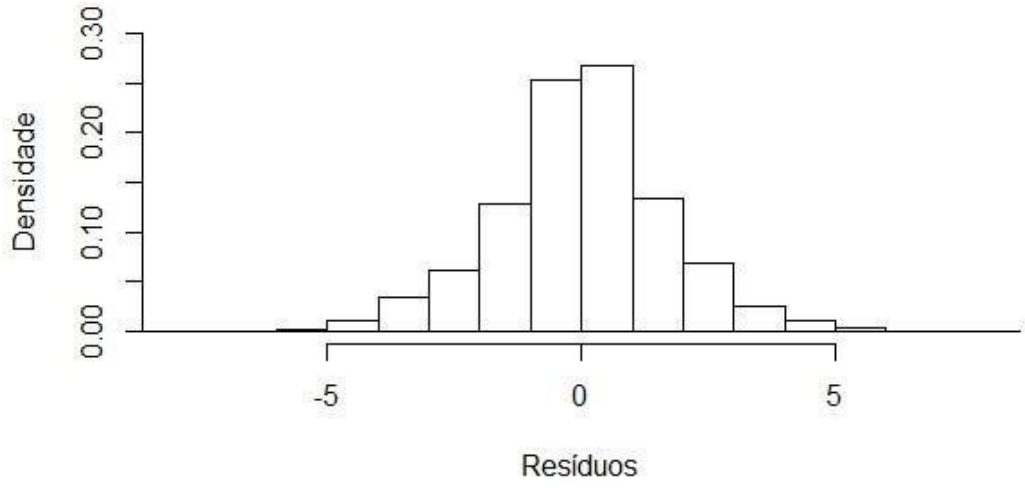


Gráfico 5: Histograma dos resíduos - 2º Ajuste Normal

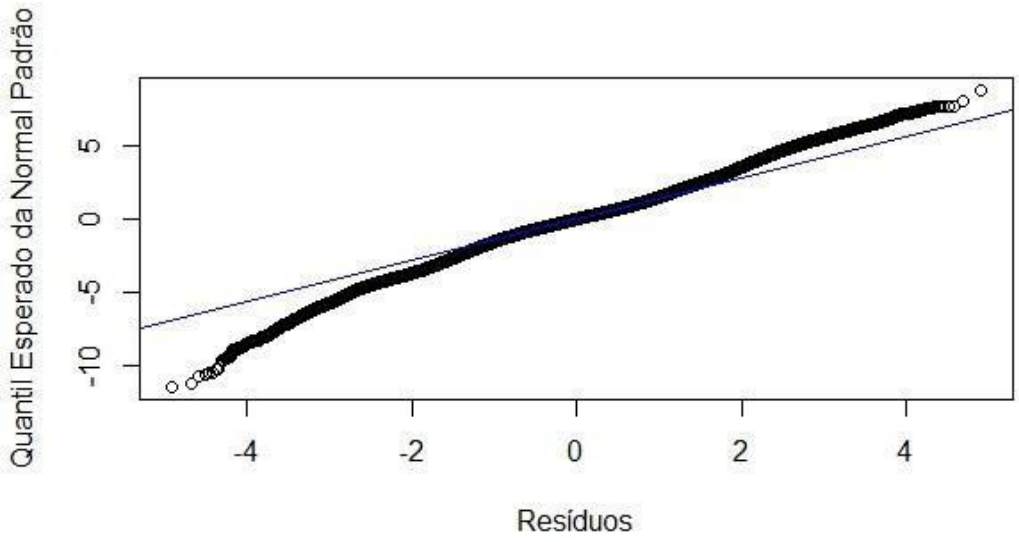


Gráfico 6: Q-Q Plot - 2º Ajuste Normal

Visto o disposto acima, a distribuição dos resíduos não pode ser normal. Portanto iremos realizar a regressão utilizando a hipótese de uma distribuição com cauda mais pesada, a t-Student.

Realizamos o ajuste do modelo adotando a hipótese de que os resíduos provêm de uma distribuição t-Student, com as mesmas variáveis e interações que se mostraram estatisticamente significativas ao final do modelo supondo normalidade. Ajustamos modelos de três até dez graus de liberdade e utilizamos o Critério de Informação de Akaike (AIC) para definir o melhor deles. Como nesse caso os parâmetros dos ajustes são os mesmos, basta compararmos o valor da log-verossimilhança entre eles, conforme mostra a tabela abaixo:

Tabela 6: Log-Verossimilhança do ajuste t-Student com diferentes graus de liberdade.

| Graus de Liberdade | Log-verossimilhança |
|---|---|
| 3 | -2.157.703 |
| 4 | -2.150.336 |
| 5 | -2.147.691 |
| 6 | -2.146.863 |
| 7 | -2.146.813 |
| 8 | -2.147.111 |
| 9 | -2.147.562 |
| 10 | -2.148.072 |

Avaliando os valores encontrados nas log-verossimilhanças dos ajustes acima, escolhemos o modelo com 7 (sete) graus de liberdade, pois ele apresenta o menor valor de AIC (nesse caso estamos avaliando a maior verossimilhança, pois os resultados são negativos). Abaixo demonstramos os valores dos coeficientes estimados.

Tabela 7: Coeficientes estimados - Ajuste t-Student.

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | 2,69000 | 0,11500 | 23,43 | $2x10^{-16}$ |
| Sexo | -0,15700 | 0,01570 | -9,96 | $2x10^{-16}$ |
| Idade | 0,01840 | 0,00022 | 84,67 | $2x10^{-16}$ |
| Tempo | -1,23000 | 0,23800 | -5,16 | $2,5x10^{-7}$ |
| $Tempo^2$ | 0,97300 | 0,17400 | 5,61 | $2,1x10^{-8}$ |
| $Tempo^3$ | -0,32800 | 0,05730 | -5,72 | $1,1x10^{-8}$ |
| $Tempo^4$ | 0,05080 | 0,00872 | 5,83 | $5,6x10^{-9}$ |
| $Tempo^5$ | -0,00292 | 0,00050 | -5,87 | $4,4x10^{-9}$ |
| Sexo:Idade | 0,00208 | 0,00032 | 6,44 | $1,2x10^{-10}$ |
| Sexo:Tempo | -0,01240 | 0,00413 | -3,00 | 0,00269 |
| Idade:Tempo | 0,00023 | 0,00006 | 4,06 | 0,00005 |
| Sexo:Idade:Tempo | 0,00031 | 0,00008 | 3,75 | 0,00018 |

Todas as variáveis e suas interações permanecem estatisticamente significativas com pelo menos 5% de nível de significância. Houve, no geral, uma redução nos valores estimados dos coeficientes, com exceção da interação entre as variáveis: sexo, idade e tempo. O p-valor também caiu, indicando uma maior significância das variáveis quando modelamos com erros t-Student.

Prosseguimos com a análise gráfica dos resíduos como mostramos a seguir:

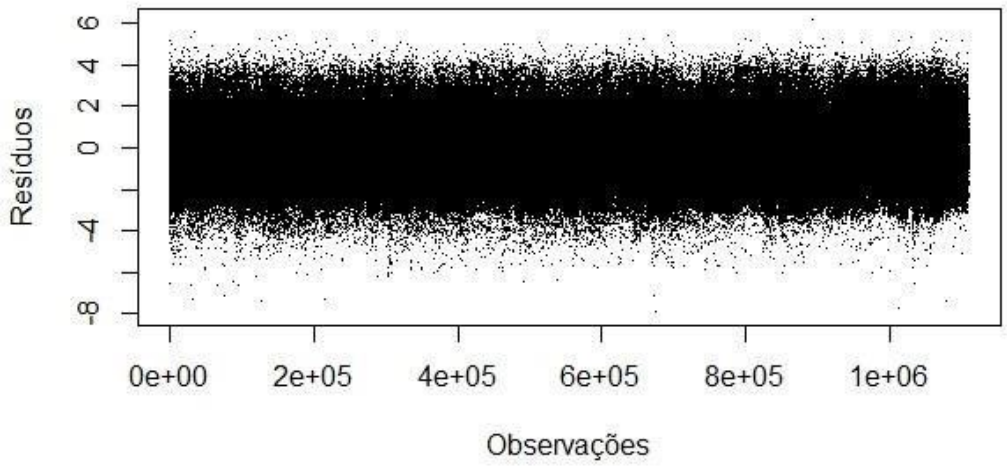


Gráfico 7: Resíduos x Observações - Ajuste t-Student

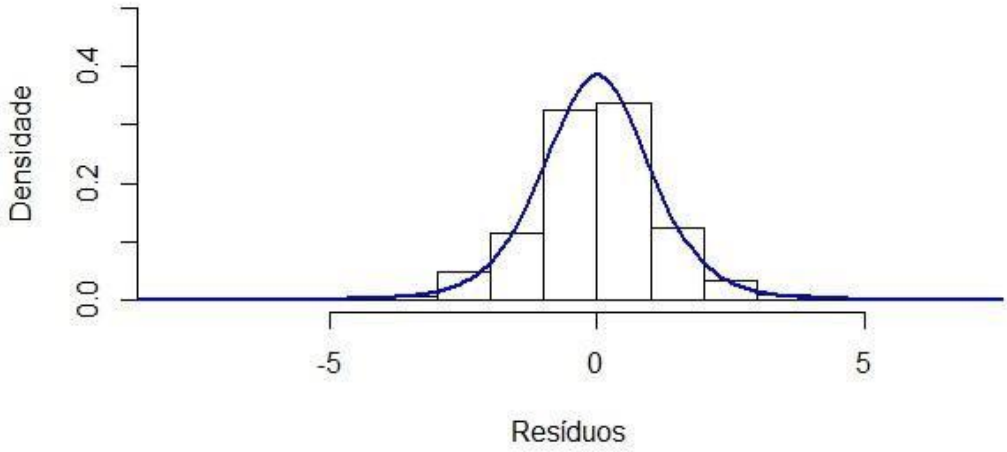


Gráfico 8: Histograma dos Resíduos - Ajuste t-Student

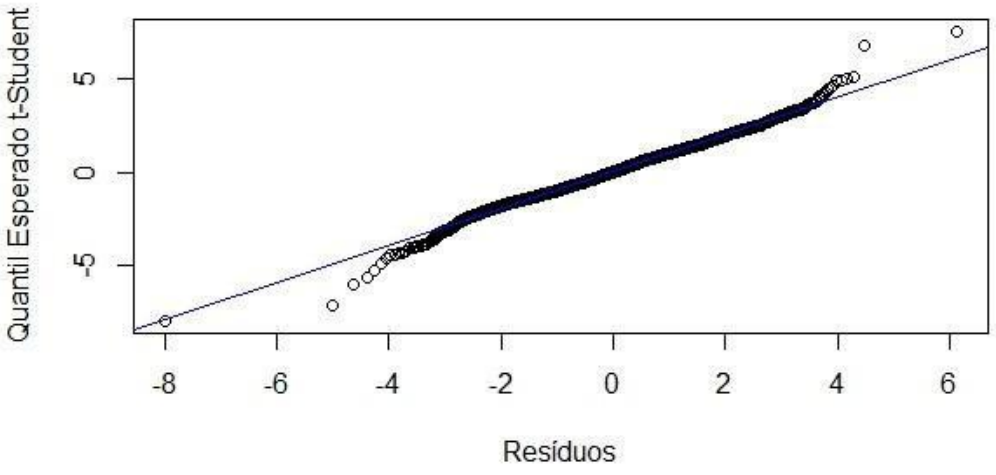


Gráfico 9: Q-Q Plot - Ajuste t-Student

Os gráficos dos resíduos se mostraram bastante satisfatórios, apesar de uma assimetria que ainda não foi modelada. Esta se deve ao fato de que alguns pontos dos dados são bastante elevados numericamente e geram uma leve assimetria à direita. Esta se reflete perfeitamente no gráfico 9, onde vemos que alguns pontos fogem da linha que demarca os quantis teóricos da distribuição t-Student.

Após diversas tentativas de uma modelagem mais detalhada para a assimetria, não conseguimos chegar a um ajuste razoável das caudas do modelo. Tentamos diversas funções do programa estatístico R, como: “cplm”, “skewt”, “sn” e também outras regressões utilizando as variáveis explicativas deste mesmo modelo para ajustarmos a assimetria dos dados. Estas funções fazem ajuste de regressão com erros normais ou t-Student assimétricos seguindo a classe de Adelchi Azzalini (1997) (“sn”), King (2004) (“skewt”) ou Zhang (2015) (“cplm”).

Porém devido à grande quantidade de dados presentes em nossa amostra tivemos diversos problemas numéricos principalmente com relação a demanda por memória física, já que todos os pacotes realizavam muitas operações e iterações, e isso levava a dificuldade de convergência das funções de maximização/minimização (nenhum deles convergiu).

Conforme dito acima, consideramos como o melhor ajuste possível para nosso objetivo neste trabalho, o modelo de regressão com erros t-Student e 7 graus de liberdade. Portanto prosseguiremos com o cálculo da previsão do valor da VCMH para o ano seguinte.

Criamos uma matriz de dados $x^0$para realizar a previsão. Os vetores incluídos na matriz foram as variáveis explicativas do modelo para poder utilizar os betas estimados pelo ajuste t-Student e obter o valor da previsão para o ano de 2014 como queríamos.

De acordo com a metodologia de previsão descrita no capítulo 2 desde trabalho, realizamos o produto vetorial entre a matriz $x^0$ e os betas estimados. Obtivemos os seguintes resultados para cada ano da variável Despesa Total Positiva:

Tabela 8: Valores da variável Despesa Total Positiva ajustada pelo modelo t-Student e seu valor original.

| Anos | Despesa Total Positiva | Despesa Total Positiva Ajustada |
|---|---|---|
| 2008 | 19.363.845,33 | 3.798.357,03 |
| 2009 | 21.585.192,05 | 4.120.405,50 |
| 2010 | 25.527.011,70 | 4.702.152,20 |
| 2011 | 30.444.306,30 | 5.276.576,99 |
| 2012 | 40.533.923,38 | 6.533.185,08 |
| 2013 | 52.038.035,36 | 8.288.827,24 |

Em todos os anos estamos subestimando a despesa total. Isso se deve ao fato de estarmos aplicando a transformação inversa da logaritimo (a exponencial). Caso o modelo fosse log-normal poderíamos utilizar um fator $\frac{\sigma^2}{2}$ de correção dentro da função exponencial. Entretanto, como este fator de correção provém de uma variável aleatória com distribuição log-normal e aqui estamos trabalhando com a distribuição t-Student, a aplicação desse fator de correção ficou comprometida, visto que a Log-t-Student não tem média ou variância bem definidas. Portanto, para corrigirmos a tendenciosidade criada pelo modelo e visando o aumento do nosso poder de previsão, faremos um procedimento *"ad hoc"*. Este será constituído por duas regressões como descritas adiante.

Primeiramente, vamos entender como se comportam as variáveis dispostas acima.

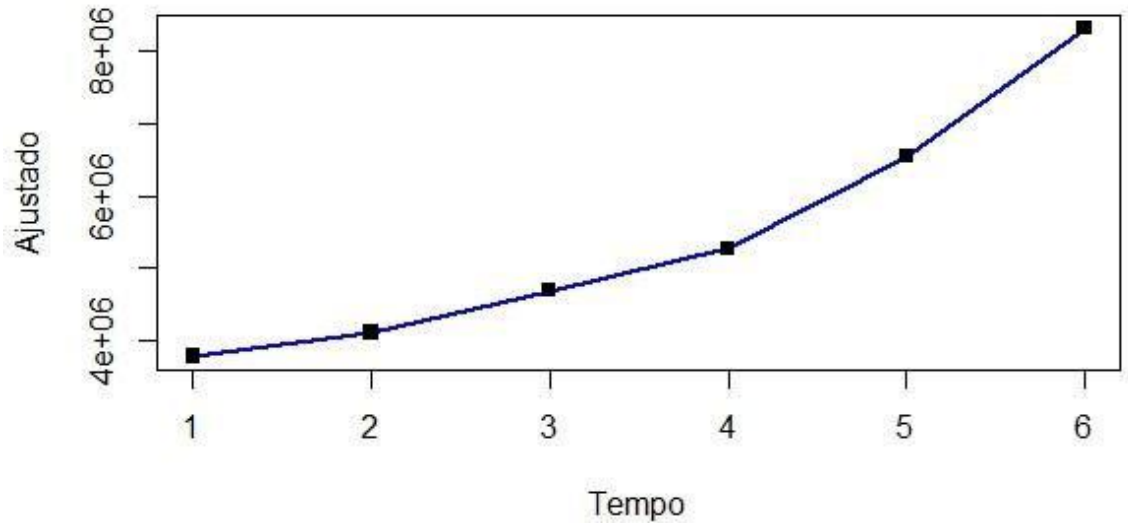


Gráfico 10: Comportamento valores ajustados

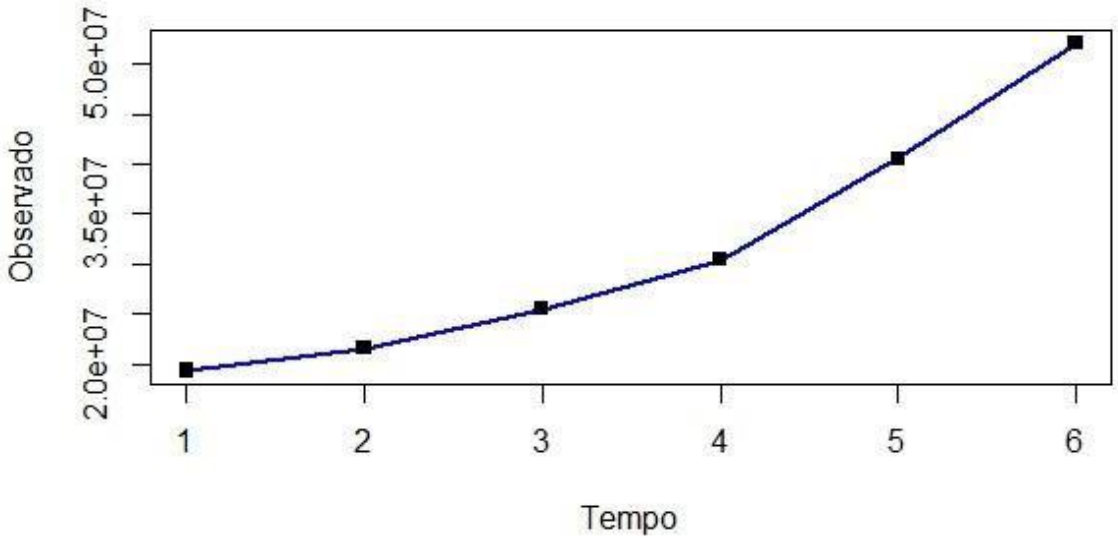


Gráfico 11: Comportamento valores observados

Graficamente, verificamos que ambas as variáveis apresentam uma tendência exponencial de crescimento ao longo do tempo. Além disso, elas apresentam 99% de correlação.

Criamos os vetores $v1$ contendo os valores de despesa total ajustados de 2008 a 2013 e o vetor $v2$ contendo os valores originais para a variável despesa total no mesmo período de $v1$.

Com isso, realizamos regressão de $v1$ sobre a função exponencial do tempo, como segue:

Modelo 1: $v1 = \beta_0 + \beta_1 t + \beta_2 e^t + \varepsilon$, sendo $\varepsilon \sim N(0, \sigma^2)$.

Tabela 9: Coeficientes estimados Modelo 1: Despesa total ajustada em função do tempo.

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | 3.251.035,0923 | 151.950,9213 | 21,3953 | 0,00022341 |
| Tempo | 45.734,0607 | 56.889,8236 | 7,9229 | 0,00419229 |
| Exp(Tempo) | 5.886,1597 | 685,0791 | 8,5919 | 0,00331446 |

Todos os coeficientes estimados são estatisticamente relevantes com pelo menos 5% de significância. Encontramos o $R^2$ igual a 99%, isso quer dizer que as variáveis independentes (tempo e a função exponencial do tempo) explicam quase por completo a

variabilidade da variável despesa total ajustada ($v1$). Isso significa que temos um poder de previsão maior para o ano de 2014.

A segunda regressão que utilizaremos para ajustar a tendenciosidade dos valores ajustados é:

Modelo 2: $v2 = \beta_0 + \beta_1 v1 + \varepsilon$ ,

sendo $\varepsilon \sim N(0, \sigma^2)$.

Tabela 10: Coeficientes estimados Modelo 2: Valores da despesa total dados em função dos ajustados.

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | -8.782.426 | 914.394,5 | -9,60464 | 0,00065687 |
| v1 | 7,40204 | 0,16133 | 45,87997 | 0,00000135 |

Assim como na regressão de $v1$ sobre o tempo, os coeficientes estimados são estatisticamente significativos ao nível de 5%. Nesse caso também encontramos 99% para o valor de $R^2$.

Calculamos a previsão do valor da despesa total ajustada para 2014 utilizando o Modelo 1. A partir daí, utilizando os coeficientes estimados pelo Modelo 2, finalmente realizamos o ajuste para o valor final da previsão da despesa total para o ano seguinte. Dessa forma, conseguimos corrigir a tendenciosidade apresentada pelo ajuste inicial t-Student.

Para nosso valor final de previsão, fizemos o intervalo de predição, como abaixo:

Tabela 11: Valor previsto 2014 e Intervalo de Predição.

| Valor Ajustado 2014 | Limite Inferior do Intervalo | Limite Superior do Intervalo |
|---|---|---|
| 86.416.143,63 | 82.626.428,04 | 90.205.859,22 |

Agora que estimamos o valor previsto para a despesa total de 2014, podemos calcular a VCMH prevista para 2014, o que é o nosso principal objetivo.

Obtemos os seguintes valores para a premissa em questão.

Tabela 12: Resultados das Previsões

| Ano | VCMH |
|---|---|
| 2009/2008 | 11,47% |
| 2010/2009 | 18,26% |
| 2011/2010 | 19,26% |
| 2012/2011 | 33,14% |
| 2013/2012 | 28,38% |
| 2014/2013 (estimado) | 66,06% |

Ao concluir nosso modelo, dado que tivemos que utilizar um procedimento "ad hoc" para corrigirmos a tendenciosidade em virtude da impossibilidade de modelagem da assimetria presente nos resíduos da regressão t-Student. Tentamos utilizar um procedimento mais simples para realizar a previsão.

Utilizamos uma regressão simples dos dados, sem tantas covariáveis ou interações destas, apenas usando o tempo e a função exponencial do tempo. O fato de escolhermos a função exponencial da variável tempo já foi justificado, pois a despesa total presente nos dados cresce desta maneira.

O ajuste utilizado foi:

$DespesaTotal = \beta_0 + \beta_1 t + \beta_2 e^t + \varepsilon$, sendo $\varepsilon \sim N(0, \sigma^2)$. E obtemos os seguintes resultados da estimação por máxima verossimilhança.

Tabela 13: Coeficientes estimados - Ajuste simplificado

| Coeficientes | Valor Estimado | Erro Padrão | Valor T | P-Valor |
|---|---|---|---|---|
| Intercepto | 14.705.536,25 | 1.744.137,56 | 8,4314 | 0,0035011 |
| Tempo | 3.601.303,03 | 652.998,2 | 5,515 | 0,0117401 |
| Exp(Tempo) | 40.261,4 | 7.863,54 | 5,12 | 0,0144217 |

Realizando a análise do p-valor encontrado, verificamos que as duas variáveis explicativas são estatisticamente significativas ao nível de 5%. O $R^2$ encontrado foi de 99%. Com os coeficientes estimados acima, realizamos a previsão para o valor da despesa total em 2014.

O valor encontrado para a premissa da VCMH que queremos estimar foi de 61,54%, ou seja, um valor muito próximo ao que havíamos encontrado anteriormente utilizando o modelo mais complexo.

## Capítulo 4: Conclusão

Ao final deste trabalho, temos algumas considerações finais a fazer. O fato de não conseguirmos modelar melhor a assimetria prejudicou um pouco o poder de previsão de nosso modelo com erros t-Student. Poderíamos ter realizado uma análise segmentada, usando as despesas agregadas por faixas etárias, respeitando a legislação ANS vigente, mas consideramos que todos os registros eram importantes e decidimos fazer da forma descrita ao longo do texto. Houve complicações numéricas e de capacidade computacional física para realizar alguns tipos de estimações, o que levou também a um maior erro quando tratamos da predição.

Utilizamos modelos de regressão e previsão bem embasados na teoria estatística e realizamos uma análise consistente dos dados e dos resultados dos ajustes. Visto isso e apesar de todas as implicações, o modelo criado foi considerado satisfatório, porém passível de um aprimoramento futuro, que não será discutido nesse trabalho.

De acordo com o apresentado neste trabalho, os resultados indicam que seria um possível modelo a ser estudado e possivelmente aplicado em termos de mercado, dependendo da companhia. Porém, a falta de sucesso em modelar a assimetria nos leva à necessidade de analisar outras maneiras de tratamento de dados para evitar esse problema. A modelagem dessa questão levaria a uma melhora de nossa capacidade de previsão, resultando assim em um resultado mais fidedigno com a realidade dos dados estudados.

**Bibliografia**

GREENE, W. H. (2012). *Econometrics Analysis (7th ed),* Pearson Education.

WEISBERG, S. (2005). *Applied Linear Regression (3rd ed),* Minnesota: Wiley.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.r-project.org/