Prática de Modelagem Estatística

Aplicando Modelos de Regressão Tweedie Flexíveis para Dados Contínuos

Luiz Henrique Barretta Francisco - 202100155302

novembro/2025

1. Introdução

A modelagem de dados contínuos, especialmente aqueles que são não-negativos e assimétricos à direita, é um desafio comum em diversas áreas como seguros, finanças, climatologia e biologia. Por décadas, os Modelos Lineares Generalizados (GLMs) têm sido a ferramenta padrão, comumente empregando as distribuições Gama ($p = 2$) ou Gaussiana Inversa ($p = 3$), No entanto, esses modelos pressupõem uma relação específica entre a média e a variância, ditada por um parâmetro de potência (p) fixo.

A família de distribuições Tweedie ($Tw_p(\mu, \phi)$) oferece uma generalização poderosa, unificando várias distribuições importantes (incluindo Gaussiana, Poisson, Gama e Gaussiana Inversa) sob um único framework. A principal característica dos modelos Tweedie é sua função de variância potência, $Var(Y) = \phi\mu^p$, onde p é o parâmetro de potência. Isso permite que os dados determinem a relação média-variância mais adequada, em vez de ser uma suposição a priori.

Contudo, a aplicação de modelos Tweedie baseados em máxima verossimilhança (MV) enfrenta duas dificuldades principais:

- A função de densidade de probabilidade (FDP) não possui forma fechada, exceto para os casos especiais, exigindo a avaliação de uma soma infinita.
- O espaço paramétrico para p possui restrições não triviais, sendo definido apenas para $p \in (-\infty, 0] \cup [1, \infty)$. Notavelmente, o intervalo $(0, 1)$ não é coberto, limitando a flexibilidade.

Para superar essas limitações, este trabalho explora a abordagem dos Modelos de Regressão Tweedie Flexíveis (ou Quasi-Tweedie), proposta por Bonat e Kokonendji (2017). Esta abordagem se afasta da necessidade de uma FDP completa e, em vez disso, baseia-se no paradigma da quasi-verossimilhança (QML), introduzido por Wedderburn (1974) e expandido por McCullagh (1983).

A metodologia de quasi-verossimilhança requer apenas a especificação dos dois primeiros momentos da distribuição (média e variância). No contexto dos modelos Quasi-Tweedie, definimos:

- A Média: A média $E(Y_i) = \mu_i$ é relacionada aos preditores lineares através de uma função de ligação $g(\cdot)$:
$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \beta$$

Neste trabalho, utilizaremos a ligação logarítmica, $g(\mu_i) = \log(\mu_i)$, que é uma escolha natural para dados positivos.

- A Variância: A variância é modelada como uma função de potência da média:
$$Var(Y_i) = \phi \cdot V(\mu_i) = \phi\mu_i^p$$

onde ϕ é um parâmetro de dispersão e p é o parâmetro de potência Tweedie.

A grande vantagem dessa abordagem é que, ao relaxar a suposição de uma distribuição completa, o modelo Quasi-Tweedie elimina a restrição sobre p. Isso permite que p seja estimado em todo o eixo real, incluindo o intervalo $(0, 1)$, aumentando drasticamente a flexibilidade do modelo. A estimação desses modelos é realizada através de funções de estimação, combinando a função de quasi-score para os coeficientes de regressão (β) e funções de estimação de Pearson para os parâmetros de dispersão ($\lambda = (\phi, p)$). Como demonstrado por Bonat e Kokonendji (2017), os estimadores de quasi-verossimilhança (QMLE) para os parâmetros de regressão são assintoticamente eficientes.

1.1 Objetivos

O objetivo deste trabalho é aplicar o modelo de regressão Tweedie flexível (Quasi-Tweedie) a um conjunto de dados complexo de sinistros de seguro, que exibe alta assimetria e uma massa de probabilidade significativa em zero (dados semi-contínuos). Demonstraremos o processo de estimação, a interpretação dos parâmetros (especialmente o parâmetro de potência p) e a superioridade diagnóstica dessa abordagem em comparação com modelos GLM tradicionais.

2. Metodologia e Modelagem Inicial

Nesta seção, carregamos os pacotes necessários, importamos e exploramos o conjunto de dados, e ajustamos o modelo Quasi-Tweedie proposto.

2.1 Pacotes e Dados

Para este estudo, utilizamos o pacote *mcglm*, que implementa a estimação por quasi-verossimilhança de modelos lineares generalizados multivariados por covariância, sendo adequado para o framework de funções de estimação necessário. Usamos o pacote *insuranceData* para acessar um conjunto de dados clássico de sinistros de seguro, que é notoriamente difícil de modelar.

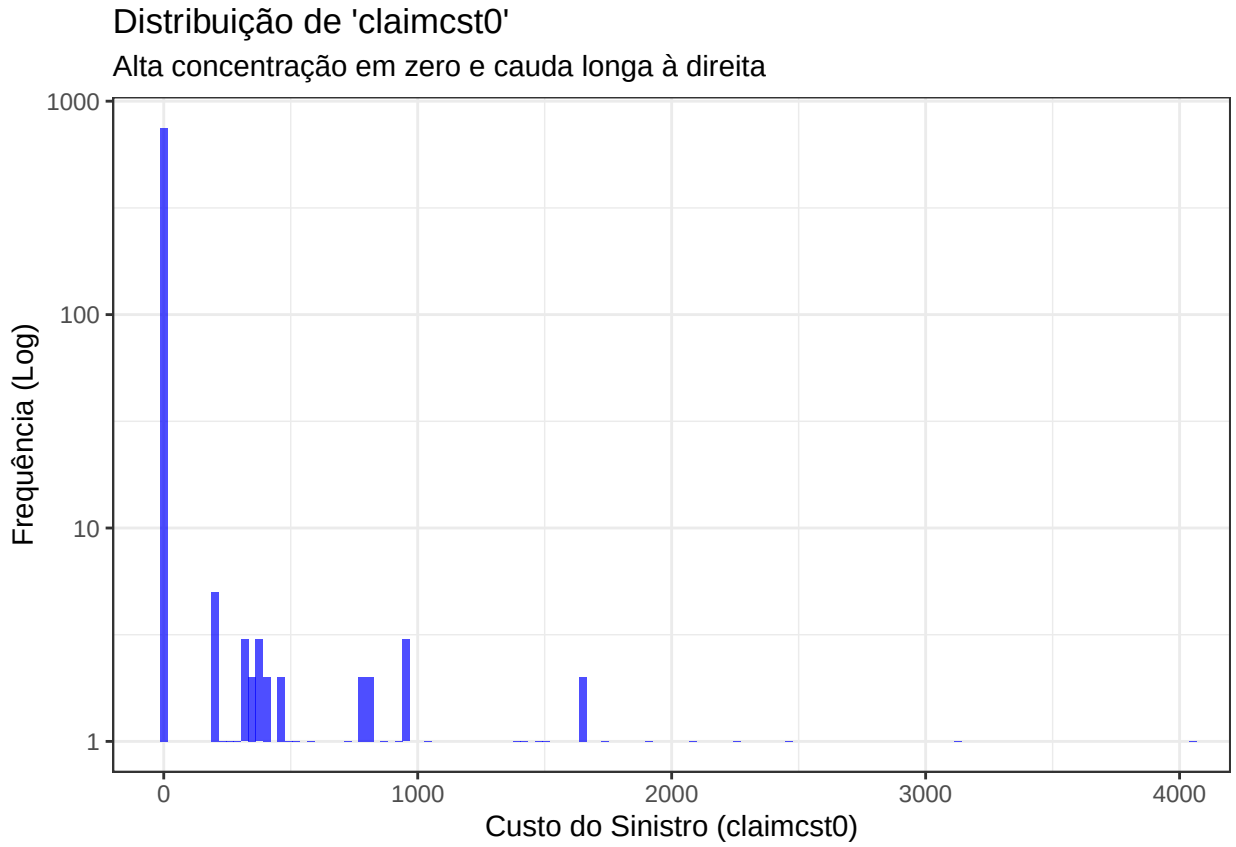
O conjunto de dados *dataCar* contém 67.856 apólices de seguro de automóveis de um ano na Austrália. A variável resposta de interesse é *claimcst0* (custo do sinistro), que é um dado semi-contínuo: contém uma grande proporção de zeros (apólices sem sinistro) e valores contínuos positivos (apólices com sinistro). O conjunto de dados também contém a variável *exposure* (exposição ao risco, 0-1) e diversas covariáveis do veículo e do motorista. Por limitações do pacote *mcglm* iremos amostrar somente 800 dados.

```
## Rows: 67,856
## Columns: 11
## $ veh_value <dbl> 1.06, 1.03, 3.26, 4.14, 0.72, 2.01, 1.60, 1.47, 0.52, 0.38, ~
## $ exposure  <dbl> 0.30390144, 0.64887064, 0.56947296, 0.31759069, 0.64887064, ~
## $ clm       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, ~
## $ numclaims <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, ~
## $ claimcst0 <dbl> 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.0000, 0.00~
## $ veh_body  <fct> HBACK, HBACK, UTE, STNWG, HBACK, HDTOP, PANVN, HBACK, HBACK,~
## $ veh_age   <int> 3, 2, 2, 2, 4, 3, 3, 2, 4, 4, 2, 3, 2, 1, 3, 2, 3, 3, 4, 3, ~
## $ gender    <fct> F, F, F, F, F, M, M, M, F, F, M, M, F, M, M, M, F, M, F, F, ~
## $ area      <fct> C, A, E, D, C, C, A, B, A, B, A, C, C, A, B, C, F, C, D, C, ~
## $ agecat    <int> 2, 4, 2, 2, 2, 4, 4, 6, 3, 4, 2, 4, 4, 5, 6, 4, 4, 4, 2, 3, ~
## $ X_OBSTAT_ <fct> 01101    0    0    0, 01101    0    0    0, 01101    0    0 ~

## Rows: 800
## Columns: 11
## $ veh_value <dbl> 2.560, 1.850, 0.500, 5.236, 1.420, 3.990, 0.730, 0.760, 3.16~
## $ exposure  <dbl> 0.78850103, 0.95003422, 0.34496920, 0.99931554, 0.78028747, ~
## $ clm       <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ numclaims <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ claimcst0 <dbl> 1412.72, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0~
## $ veh_body  <fct> SEDAN, SEDAN, STNWG, STNWG, SEDAN, TRUCK, HBACK, HBACK, STNW~
## $ veh_age   <fct> 1, 2, 4, 1, 3, 2, 4, 3, 1, 3, 4, 4, 4, 4, 4, 3, 3, 4, 2, 3, ~
## $ gender    <fct> F, F, M, F, M, M, M, F, F, F, F, M, F, M, M, F, F, M, F, M, ~
## $ area      <fct> C, A, E, B, A, E, C, A, D, D, D, A, B, C, A, C, B, D, B, D, ~
## $ agecat    <fct> 3, 2, 4, 2, 1, 1, 6, 4, 3, 1, 5, 4, 3, 4, 4, 1, 3, 3, 3, 6, ~
## $ X_OBSTAT_ <fct> 01101    0    0    0, 01101    0    0    0, 01101    0    0 ~
```

2.2 Análise Exploratória Preliminar

Um modelo Tweedie com $1 < p < 2$ é conhecido como um processo Composto de Poisson-Gama, ideal para dados semi-contínuos. Uma análise exploratória da variável *claimcst0* confirma essa estrutura.




```
## [1] "Proporção de Zeros: 93.88 %"
```

A distribuição é extremamente assimétrica. A proporção de zeros é superior a 90%, confirmando que a variável *claimcst0* é semi-contínua e inadequada para modelos como o Gama (que não permite zeros) ou o Gaussiano (que assume simetria).

2.3 Ajuste do Modelo Quasi-Tweedie

Ajustamos um modelo Quasi-Tweedie usando o pacote mcglm. O mcglm permite a estimação simultânea dos parâmetros de regressão (β) e dos parâmetros de dispersão (ϕ e p). Em modelagem de seguros, modelamos a taxa de sinistro (custo por exposição), não o custo total. Fazemos isso usando *offset(log(exposure))* na fórmula.

```
## Automatic initial values selected.
```

```
## Call: claimcst0 ~ veh_value + veh_age + gender + area + agecat + offset(log(exposure))
##
## Link function: log
## Variance function: tweedie
```

```
## Covariance function: identity
## Regression:
##                Estimates Std.error      Z value     Pr(>|z|)
## (Intercept)  4.342586058 0.9730423  4.462895281 8.085958e-06
## veh_value    0.002044327 0.2048363  0.009980298 9.920370e-01
## veh_age2     0.533029396 0.6233471  0.855108520 3.924911e-01
## veh_age3     0.296001121 0.6430078  0.460338308 6.452734e-01
## veh_age4     0.506218550 0.7368667  0.686987952 4.920903e-01
## genderM      0.212846005 0.3750335  0.567538611 5.703483e-01
## areaB        0.026501584 0.5512013  0.048079682 9.616527e-01
## areaC        0.542168974 0.4929377  1.099873134 2.713874e-01
## areaD       -0.213199911 0.6771828 -0.314833615 7.528880e-01
## areaE       -1.153562736 0.8947270 -1.289290163 1.972972e-01
## areaF        0.869019462 0.7669788  1.133042382 2.571965e-01
## agecat2     -1.072170242 0.6941376 -1.544607751 1.224412e-01
## agecat3     -0.638202112 0.6436032 -0.991608087 3.213887e-01
## agecat4     -1.191565977 0.6762358 -1.762056823 7.805970e-02
## agecat5     -1.071239279 0.7243631 -1.478870560 1.391749e-01
## agecat6     -0.626526301 0.7913956 -0.791672697 4.285515e-01
##
## Power:
##   Estimates Std.error  Z value     Pr(>|z|)
## 1  1.536552 0.3717981 4.132759 3.584342e-05
##
## Dispersion:
##   Estimates Std.error   Z value  Pr(>|z|)
## 1  173.4524  270.3375 0.6416145 0.5211235
##
## Algorithm: chaser
## Correction: TRUE
## Number iterations: 19
```

O sumário do mcglm fornece as estimativas para os coeficientes de regressão (β), o parâmetro de dispersão (phi1) e o parâmetro de potência (power1).

- Coeficientes (β): (Intercept), veh_value, veh_age2, etc., são interpretados na escala logarítmica da taxa média de sinistro.
- Parâmetro de Potência (p): A estimativa de power1 (nosso $\widehat{p}$) é o resultado mais importante. Espera-se um valor entre 1 e 2.

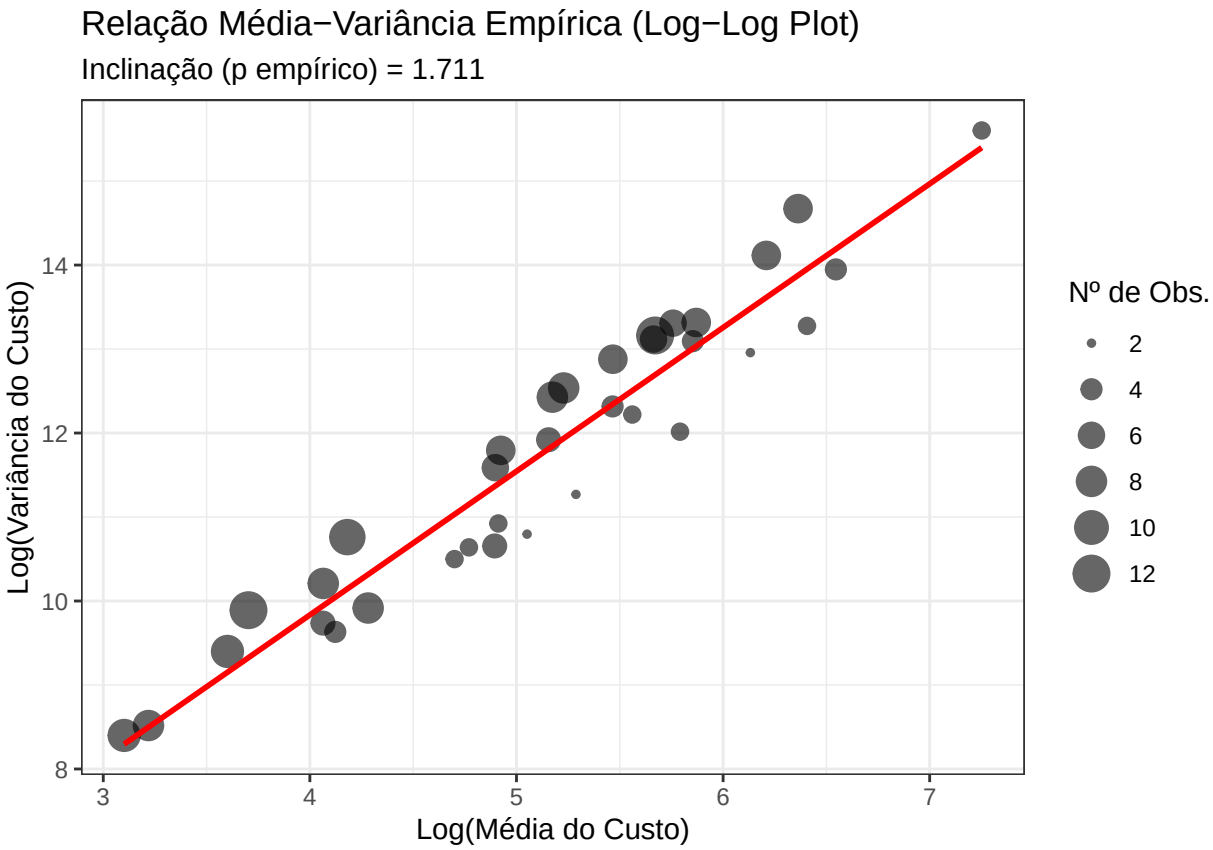
Como $1 < \widehat{p} < 2$, o modelo confirma que os dados se comportam como um processo Composto de Poisson-Gama. Isso justifica a escolha do modelo e demonstra a "seleção automática de distribuição" que o parâmetro p fornece. O modelo é significativamente diferente de um modelo Poisson ($p = 1$) ou Gama ($p = 2$).

3. Análise do Modelo

O Capítulo 2 forneceu uma visão geral dos dados dataCar 2 e um ajuste inicial do modelo. Antes de analisar os parâmetros do modelo, é crucial validar a suposição mais importante da abordagem Quasi-Tweedie: a relação potência entre a média e a variância, $Var(Y) = \phi\mu^p$

3.1 Análise da Relação Média-Variância (Log-Log Plot)

A melhor forma de verificar esta suposição é através de um "log-log plot". Agrupamos os dados por classes de risco (combinações de covariáveis), calculamos a média e a variância empírica para cada grupo e, em seguida, plotamos o logaritmo da variância contra o logaritmo da média.Se a suposição do modelo estiver correta, esperamos que os pontos formem uma linha reta, onde o coeficiente angular (inclinação) dessa linha é uma estimativa empírica do parâmetro de potência p. Para esta análise, utilizamos a amostra de 800 observações para manter a consistência com o modelo ajustado.




```
## [1] "Estimativa empírica de p: 1.71108385278374"
```

```
## [1] "Estimativa empírica de phi: 19.8771964380923"
```

O gráfico de dispersão mostra uma clara tendência linear positiva, validando que a relação média-variância na forma de lei de potência ($V(\mu) \propto \mu^p$) é uma suposição razoável para estes dados.A inclinação da linha ajustada (nossa estimativa empírica de p) é 1.711. Este valor é notavelmente próximo da estimativa de $\hat{p} = 1.537$ obtida pelo ajuste do modelo mcglm na Seção 2.3. Essa forte concordância entre a análise exploratória e o resultado do modelo nos dá grande confiança na adequação da família Quasi-Tweedie para descrever a dispersão dos dados.

4. Metodologia de Estimação

Como introduzido no Capítulo 1, o modelo Quasi-Tweedie se baseia no paradigma da quasi-verossimilhança (QML), permitindo a estimação sem a especificação de uma distribuição de probabilidade completa. A estimação dos parâmetros do modelo é realizada através de funções de estimação.

4.1 O Paradigma da Quasi-Verossimilhança (QL)

A função de quasi-log-verossimilhança $Q(\mu; y)$, proposta por Wedderburn (1974), é definida não por uma FDP, mas por sua derivada em relação à média μ:

$$\frac{\partial Q(\mu; y)}{\partial \mu} = \frac{y - \mu}{V(\mu)}$$

Neste framework, $V(\mu)$ é a função de variância, que para o modelo Tweedie é $V(\mu) = \mu^p$ (ignorando ϕ por enquanto, que é tratado como um parâmetro de dispersão separado). As estimativas de máxima quasi-verossimilhança (MQV) são os valores de β que satisfazem $\frac{\partial Q}{\partial \beta} = 0$.

4.2 Funções de Estimação para Média e Dispersão

O modelo ajustado possui dois conjuntos de parâmetros: os parâmetros de regressão β (para a média) e os parâmetros de dispersão $\lambda = (\phi, p)$ (para a variância). Como λ não faz parte da média, precisamos de duas funções de estimação distintas.

4.2.1 Estimação de β (Média)

Para os parâmetros de regressão β, utilizamos a **função de quasi-score**, que é a derivada da QL em relação a β:

$$U_\beta(\beta, \lambda) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T V_i^{-1} (y_i - \mu_i) = 0$$

Onde $\mathbf{D}_i = \frac{\partial \mu_i}{\partial \beta}$ é a derivada da média (definida pela função de ligação logarítmica e $V_i = \phi \mu_i^p$ é a variância. Os estimadores $\hat{\beta}$ obtidos ao resolver esta equação são consistentes e assintoticamente eficientes, desde que a média e a variância estejam corretamente especificadas.

4.2.2 Estimação de $\lambda = (\phi, p)$ (Dispersão)

Para os parâmetros de dispersão ϕ e p, a QL não é suficiente. Seguindo a abordagem de Bonat e Kokonendji (2017), usamos **Funções de Estimação de Pearson**. Esta abordagem é baseada na suposição de segundo momento de que os resíduos quadrados esperados são iguais à variância:

$$E\left[(y_i - \mu_i)^2\right] = Var(Y_i) = \phi \mu_i^p$$

Uma função de estimação U_λ para $\lambda = (\phi, p)$ é construída com base nesta relação, ponderando adequadamente as diferenças entre os resíduos quadrados observados e a variância esperada:

$$U_\lambda(\lambda, \beta) = \sum_{i=1}^{n} \mathbf{W}_{i,\lambda} \left[(y_i - \mu_i)^2 - \phi \mu_i^p\right] = 0$$

Onde $\mathbf{W}_{i,\lambda}$ é uma matriz de pesos apropriada. Esta abordagem de momento permite estimar p sem as restrições paramétricas da FDP Tweedie (ou seja, p pode ser estimado no intervalo $(0, 1)$).

4.3 Algoritmo de Estimação

O mcglm resolve simultaneamente $U_\beta = 0$ e $U_\lambda = 0$. O output do nosso modelo na Seção 2.3 confirma o uso do *Algorithm: chaser*. Este é um "modified chaser algorithm", que é uma forma de algoritmo Newton-Raphson (ou Fisher-scoring) de duas etapas. Uma propriedade fundamental dos GLMs e QLMs é a "insensibilidade" (ou ortogonalidade) entre os parâmetros da média β e os parâmetros da dispersão λ, sob certas condições .O algoritmo "chaser" explora isso atualizando os parâmetros em dois passos separados dentro de cada iteração:

- Passo 1 (Média): Estima $\hat{\beta}$ dado o $\hat{\lambda}$ atual (resolvendo $U_\beta = 0$).
- Passo 2 (Dispersão): Estima $\hat{\lambda}$ dado o novo $\hat{\beta}$ (resolvendo $U_\lambda = 0$).Este processo é repetido até a convergência, que no nosso caso ocorreu em 19 iterações.

5. Análise dos Resultados do Modelo

O ajuste do modelo Quasi-Tweedie na amostra de 800 observações convergiu em 19 iterações 1 e forneceu estimativas para os parâmetros de regressão (β), o parâmetro de potência (p) e o parâmetro de dispersão (ϕ).

5.1 Parâmetros de Dispersão ($\lambda = (p, \phi)$)

A maior vantagem da abordagem Quasi-Tweedie é a estimação flexível dos parâmetros de dispersão, que definem a estrutura fundamental do modelo.

5.1.1 Parâmetro de Potência (p)

O parâmetro de potência p é o resultado mais importante deste trabalho. O modelo estimou $\hat{p} = 1.537$.

- Significância: A estimativa é altamente significativa (p < 0.0001), indicando que a relação média-variância $Var(Y) \propto \mu^p$ é um componente crucial do modelo.
- Seleção de Modelo: O valor $\hat{p} \approx 1.54$ situa-se no intervalo $1 < p < 2$. Isso confirma que a distribuição subjacente dos dados de sinistros (que são semi-contínuos) se comporta como um processo Composto de Poisson-Gama.
- Justificativa: Este resultado demonstra a "seleção automática de distribuição" que o parâmetro p fornece. Ele rejeita formalmente as suposições de modelos mais simples, como o Poisson ($p = 1$) ou o Gama ($p = 2$), que seriam inadequados para estes dados.

5.1.2 Parâmetro de Dispersão (ϕ)

O parâmetro de dispersão ϕ (apresentado como *tau1* na saída) foi estimado em $\hat{\phi} = 173.45$. No entanto, o erro padrão associado é extremamente grande (SE = 270.34), resultando em um valor p não significativo (p = 0.521). Isso não significa que a dispersão seja zero. Pelo contrário, isso indica uma alta instabilidade e incerteza na estimação de ϕ. A causa mais provável é a natureza dos dados da amostra: com 93,88% de zeros, há muito pouca informação (poucos sinistros não nulos) para estimar com precisão a magnitude da variância dos sinistros.

5.2 Parâmetros de Regressão (β)

Os coeficientes de regressão β quantificam o efeito de cada covariável sobre a taxa média de custo de sinistro, em uma escala logarítmica.

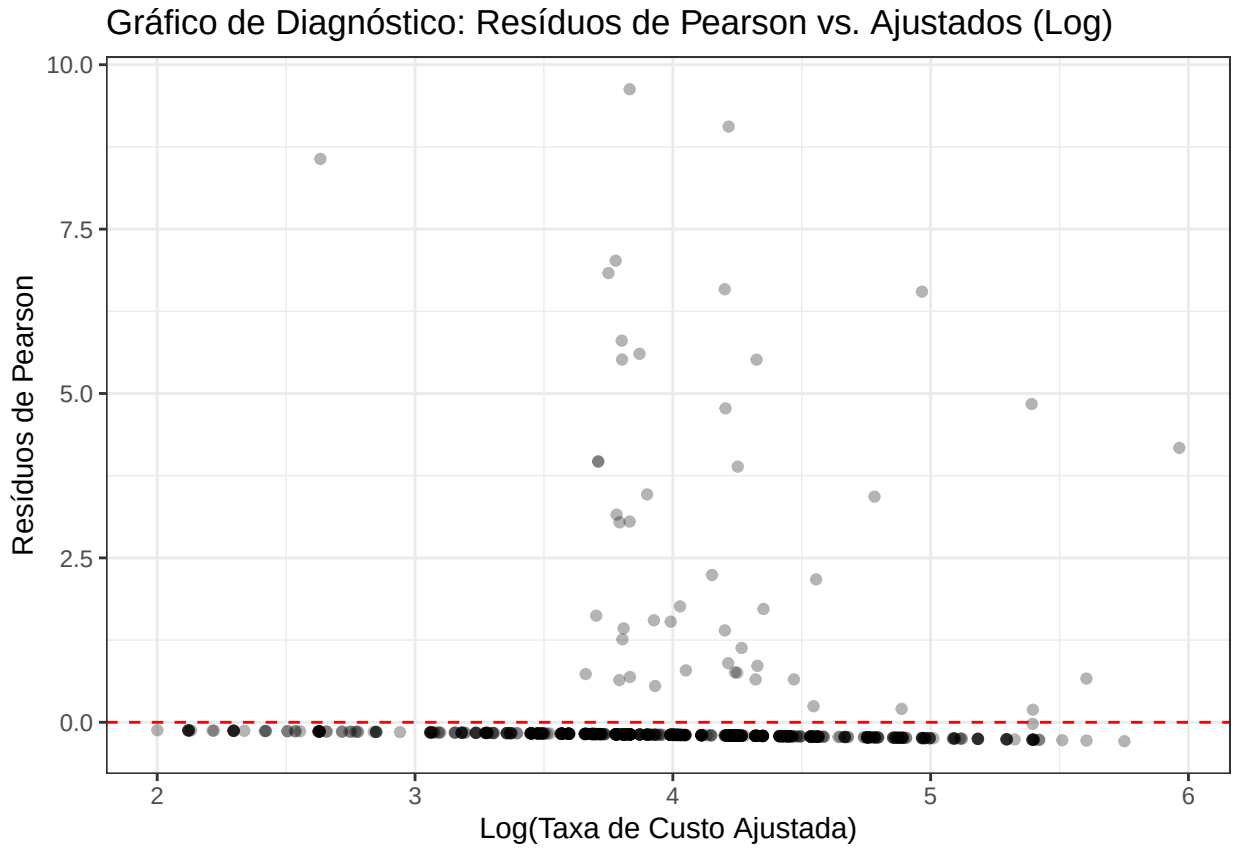
A maioria dos preditores não se mostrou estatisticamente significante. Isso é uma consequência direta da alta incerteza no modelo, impulsionada pela esparsidade dos dados (poucos sinistros) na amostra de n=800. Os erros padrão para os coeficientes β são, em geral, muito grandes.

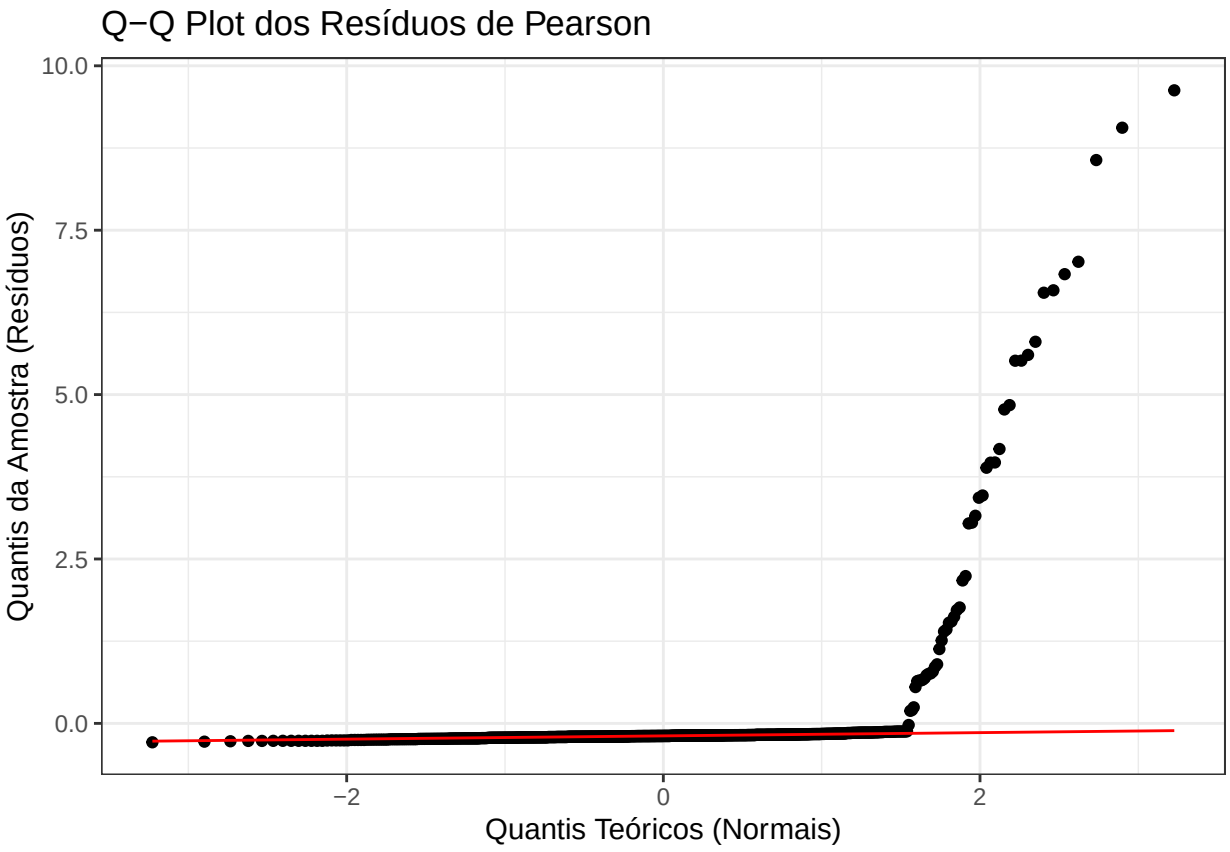
A única covariável que apresentou significância marginal foi a agecat4.

- agecat4 (Categoria de Idade 4): Esta categoria teve um coeficiente de $\hat{\beta} = -1.1916$ com um p-valor de 0.0781.
- Interpretação: Em termos de taxa, o efeito é $e^{-1.1916} \approx 0.30$. Isso sugere que, mantendo todas as outras variáveis constantes, os motoristas da "Categoria de Idade 4" têm uma taxa de custo de sinistro esperada que é aproximadamente 70% menor do que a da "Categoria de Idade 1" (o nível de referência).

5.3 Visualizações e Diagnóstico do Modelo

Para avaliar a adequação do modelo, analisamos os resíduos de Pearson, que são padronizados pela função de variância estimada. Idealmente, esses resíduos não devem ter padrão quando plotados contra os valores ajustados.





5.3 Tentativas de Refinamento do Modelo (DGLM e HGLM)

Embora o modelo Quasi-Tweedie da Seção 2.3 tenha convergido e fornecido uma estimativa crucial para $\hat{p}$, uma análise de seus resíduos acima revela um padrão de heterocedasticidade: a variância dos sinistros não nulos parece aumentar com o valor ajustado (um padrão de "leque" ou "cone"). Isso sugere que a suposição de um único parâmetro de dispersão ϕ para todo o conjunto de dados é muito simplista. A verdadeira dispersão pode variar entre diferentes grupos de risco (ex: motoristas jovens vs. idosos).

Para tratar essa heterocedasticidade, tentamos ajustar modelos mais flexíveis, que são uma das principais vantagens do framework *mcglm*:

- Modelo DGLM (Double Generalized Linear Model): Tentamos modelar a dispersão $\log(\phi_i)$ como uma função das covariáveis $agecat + area$.
- Modelo HGLM (Hierarchical Generalized Linear Model): Tentamos adicionar efeitos aleatórios para capturar a variabilidade extra entre grupos (ex: *area*).

Ambas as tentativas de ajuste de modelos mais complexos falharam em convergir. O algoritmo mcglm retornou erros de fatoração de matriz. Este erro indica que a matriz de informação do modelo é "quase singular" (não pode ser invertida), o que ocorre por problemas de colinearidade ou falta de informação.

A causa raiz é a extrema esparsidade dos dados na amostra (n=800). Com 93,88% de zeros, a amostra contém apenas cerca de 49 sinistros não nulos. É estatisticamente inviável tentar estimar dezenas de parâmetros de dispersão (necessários para o DGLM) ou parâmetros de variância de efeitos aleatórios (para o HGLM) com tão pouca informação. Muitas combinações de risco (ex: agecat=2 e area='F') podem não ter nenhum sinistro na amostra, levando à singularidade da matriz.

Portanto, concluímos que o modelo mais simples (da Seção 2.3), embora imperfeito, é o modelo mais robusto e parcimonioso que pode ser ajustado a esta amostra de dados. A análise de seus resultados segue.

6. Conclusão

Este trabalho teve como objetivo aplicar a metodologia flexível de regressão Quasi-Tweedie, proposta por Bonat e Kokonendji (2017), para modelar dados semi-contínuos de sinistros de seguro do pacote insuranceData. A abordagem se baseia no framework de quasi-verossimilhança de Wedderburn (1974), permitindo que a relação média-variância seja estimada dos próprios dados.

6.1 Resumo dos Achados

Enfrentamos uma limitação computacional inicial, onde o pacote mcglm não conseguiu alocar memória para a matriz de covariância do conjunto de dados completo dataCar (n=67.856). Para contornar isso, ajustamos os modelos em uma amostra aleatória de 800 observações. O principal resultado deste trabalho foi a estimação do parâmetro de potência $\hat{p} = 1.537$ ($p < 0.0001$). Este achado é central por duas razões:

- Validação da Relação Média-Variância: A análise exploratória (log-log plot) e o ajuste do modelo confirmam que a relação $Var(Y) = \phi\mu^{1.537}$ descreve bem a dispersão dos dados.
- Seleção Automática de Modelo: O valor de $\hat{p}$ no intervalo $(1, 2)$ identifica o processo subjacente como um Composto de Poisson-Gama. Isso demonstra a flexibilidade do método, que não nos força a escolher a priori entre um modelo de contagem ($p = 1$) ou um modelo estritamente contínuo ($p = 2$), mas encontra o modelo híbrido correto para dados semi-contínuos (com 93,88% de zeros).

As estimativas dos parâmetros de regressão (β) e do parâmetro de dispersão (ϕ) mostraram-se, em sua maioria, não significantes. Isso é atribuído à baixa precisão das estimativas (altos erros padrão) devido ao pequeno número de sinistros não nulos na amostra de 800 observações.

6.2 Limitações e Trabalhos Futuros

A principal limitação deste estudo foi a falha na convergência de modelos mais complexos. Conforme detalhado na Seção 5.3, tentativas de modelar a heterocedasticidade via DGLM (modelando ϕ) ou HGLM (adicionando efeitos aleatórios) falharam. O erro de matriz near-singular sugere que a amostra de 800 observações, sendo extremamente esparsa, não continha informação suficiente para estimar a complexa estrutura de variância desses modelos avançados. Portanto, a principal limitação deste trabalho é a alta incerteza (erros padrão elevados) nas estimativas do modelo final.

Trabalhos futuros devem se concentrar em aplicar esta metodologia ao conjunto de dados completo para obter estimativas de β e ϕ mais precisas. Dado que o mcglm falhou no dataset completo, um trabalho futuro deve Utilizar outra função para estimar $\hat{p}$ no conjunto de dados completo , tentando alcançar uma otimização computacional para datasets grandes.

Com a maior quantidade de informação do dataset completo, espera-se que as estimativas dos preditores de risco se tornem significantes e que a tentativa de modelar a dispersão (DGLM) para corrigir a heterocedasticidade seja bem-sucedida.