# Você sobreviveria ao Titanic?

Eu, Vítor Pereira, não.

## Sumário



# 1 Introdução

Neste trabalho iremos relembrar uma das maiores tragédias da história, Titanic, em 1912, provavelmente o desastre marítimo mais conhecidos da história. Assim, por meio desse trabalho pretendemos simular situações em que pessoas da atualidade fossem transportadas para a época do desastre, elas sobreviveriam ao Titanic? Utilizando regressão logística podemos prever a chance que qualquer pessoa teria de sobreviver à tragédia, assim utilizaremos alguns modelos para tentar garantir o máximo de confiabilidade possível.

# 2 Modelagem

Começaremos com dois modelos de regressão logística, um modelo utilizando a engine `glm`, muito conhecida e utilizada para realização de modelos inferenciais e a engine `glmnet` muito utilizada para modelos preditivos de aprendizado de máquina.

## 2.1 Dados de teste e treino

Realizamos a divisão da base de dados completa em outras duas: Dados de treino e Dados de teste, para assim pode verificar se os modelos propostos são bons para previsão fora da amostra, sem problemas de **overfitting** e **underfitting**. Assim obtivemos as seguintes medidas para os dados de teste utilizando os modelos ajustados com os dados de treino:

Tabela 1: Estatísticas do Modelo 1 ajustado com os dados de treino

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -1.295 | 0.277 | -4.67 | 0 |
| Age | -0.036 | 0.009 | -4.21 | 0 |
| Pclass_X1 | 2.363 | 0.310 | 7.61 | 0 |
| Pclass_X2 | 1.145 | 0.279 | 4.11 | 0 |
| Sex_female | 2.430 | 0.234 | 10.40 | 0 |

Percebemos que esse modelo é um modelo de regressão logística comum, em que conseguimos obter erro padrão, estatística e p-valores, podendo fazer uma robusta análise inferencial.

Tabela 2: Estatísticas do Modelo 2 ajustado com os dados de treino

| term | estimate | penalty |
|---|---|---|
| (Intercept) | -1.197 | 0 |
| Age | -0.025 | 0 |
| Pclass_X1 | 1.752 | 0 |
| Pclass_X2 | 0.813 | 0 |
| Sex_female | 2.045 | 0 |

Ao contrário do modelo de cima, esse é um modelo focado para previsão dos dados, envolvendo mais técnicas de aprendizado de máquina, assim não podemos realizar a análise inferencial.

Tabela 3: Métricas de Avaliação do Modelo 1 nos Dados de Treino

| .metric | .estimate |
|---|---|
| accuracy | 0.821 |
| kap | 0.635 |
| precision | 0.806 |
| sens | 0.888 |
| spec | 0.741 |

Tabela 4: Métricas de Avaliação do Modelo 2 nos Dados de Treino

| .metric | .estimate |
|---|---|
| accuracy | 0.810 |
| kap | 0.611 |
| precision | 0.786 |
| sens | 0.898 |
| spec | 0.704 |

Percebemos que as métricas nos dois Modelos são muito próximas, porém o Modelo 1 que utiliza a engine `glm` padrão, tem melhores valores em Precisão e Especificidade e o Kappa de Cohen. O Modelo 2 tem Sensibilidade superior, mas quanto a Acurácia os dois modelos empatam. No entanto, são boas métricas para ambos os modelos assim seguiremos com eles, porém agora unindo os dados de treino e de teste.

## 2.2 Reajuste aos dados completos

Como a previsão para os dados de teste está boa, podemos reajustar e usar o banco de dados completo.

Tabela 5: Estatísticas do Modelo 1

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -1.326 | 0.248 | -5.35 | 0 |
| Age | -0.037 | 0.008 | -4.83 | 0 |
| Pclass_X1 | 2.581 | 0.281 | 9.17 | 0 |
| Pclass_X2 | 1.271 | 0.244 | 5.21 | 0 |
| Sex_female | 2.523 | 0.207 | 12.16 | 0 |

Ajustando o Modelo 1, com todos os dados ainda obtemos significância em todas as variáveis.

Tabela 6: Métricas de Avaliação do Modelo 1

| .metric | .estimate |
|---|---|
| accuracy | 0.789 |
| kap | 0.558 |
| precision | 0.811 |
| sens | 0.840 |
| spec | 0.714 |

Em relação as métricas, acontece algo curioso em relação ao modelo ajustado com os dados de teste, em que apenas a Sensibilidade acaba aumentando, mas ainda são bons valores, a Acurácia é de 78,9%.

Tabela 7: Estatísticas do Modelo 2

| term | estimate | penalty |
|---|---|---|
| (Intercept) | -1.210 | 0 |
| Age | -0.024 | 0 |
| Pclass_X1 | 1.858 | 0 |
| Pclass_X2 | 0.870 | 0 |
| Sex_female | 2.097 | 0 |

Ajustando o Modelo 2, possuímos valores extremamente semelhantes para as estimativas dos $\beta's$, assim as previsões devem permanecer parecidas.

Tabela 8: Métricas de Avaliação do Modelo 2

| .metric | .estimate |
|---|---|
| accuracy | 0.789 |
| kap | 0.558 |
| precision | 0.811 |
| sens | 0.840 |
| spec | 0.714 |

Para o modelo 2, aconteceu uma conjuntura semelhante ao modelo 1, em que as métricas de avaliação acabam diminuindo do Modelo com os Dados de Treino para o Modelo com todos os dados, porém nesse caso nem a Sensibilidade aumentou. No entanto, analisando os valores percebemos que eles são semelhantes aos valores do Modelo 1.

### 2.3 Equações dos modelos

,

$$, \log\left[\frac{P(..\text{y}=1)}{1-P(..\text{y}=1)}\right] = \alpha + \beta_1(\text{Age}) + \beta_2(\text{Pclass\_X1}) + \beta_3(\text{Pclass\_X2}) + \beta_4(\text{Sex\_female}) , \quad (1)$$

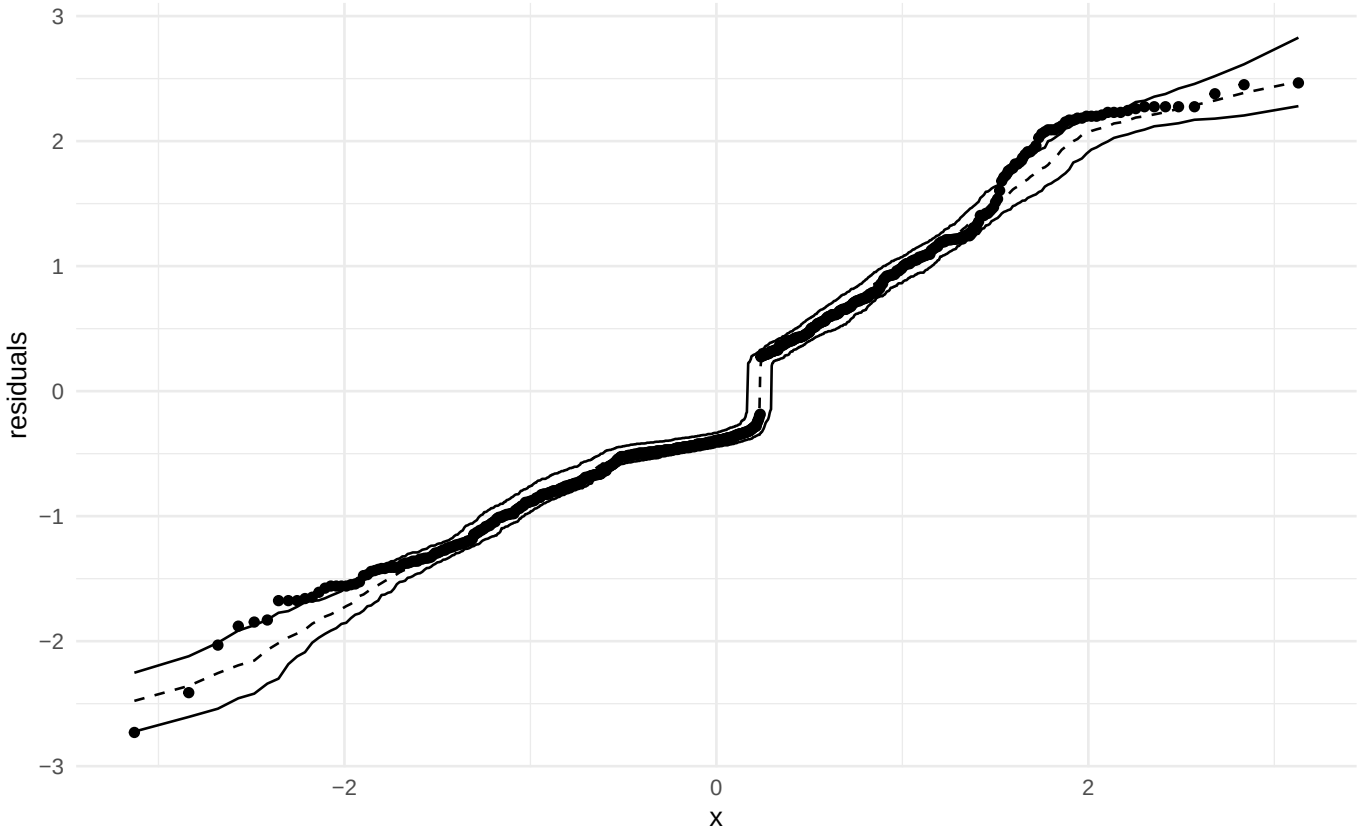
## 3 Análise de Influência

Nesta seção será realizada uma busca de observações atípicas no banco de dados, que assim possam estar influenciado a análise, também influenciado pelas junções de tipos realizados anteriomente, assim utilizaremos 5 análises para a verificação de pontos de influência: Análise de Resíduos Deviance, Envelope Simulado, Distância de Cook, Alavancagem e DFFits. No entanto, não foi possível realizar essas analises para o Modelo 2 pois ele possui a engine `glmnet`.

### 3.1 Resíduos deviances vs indices

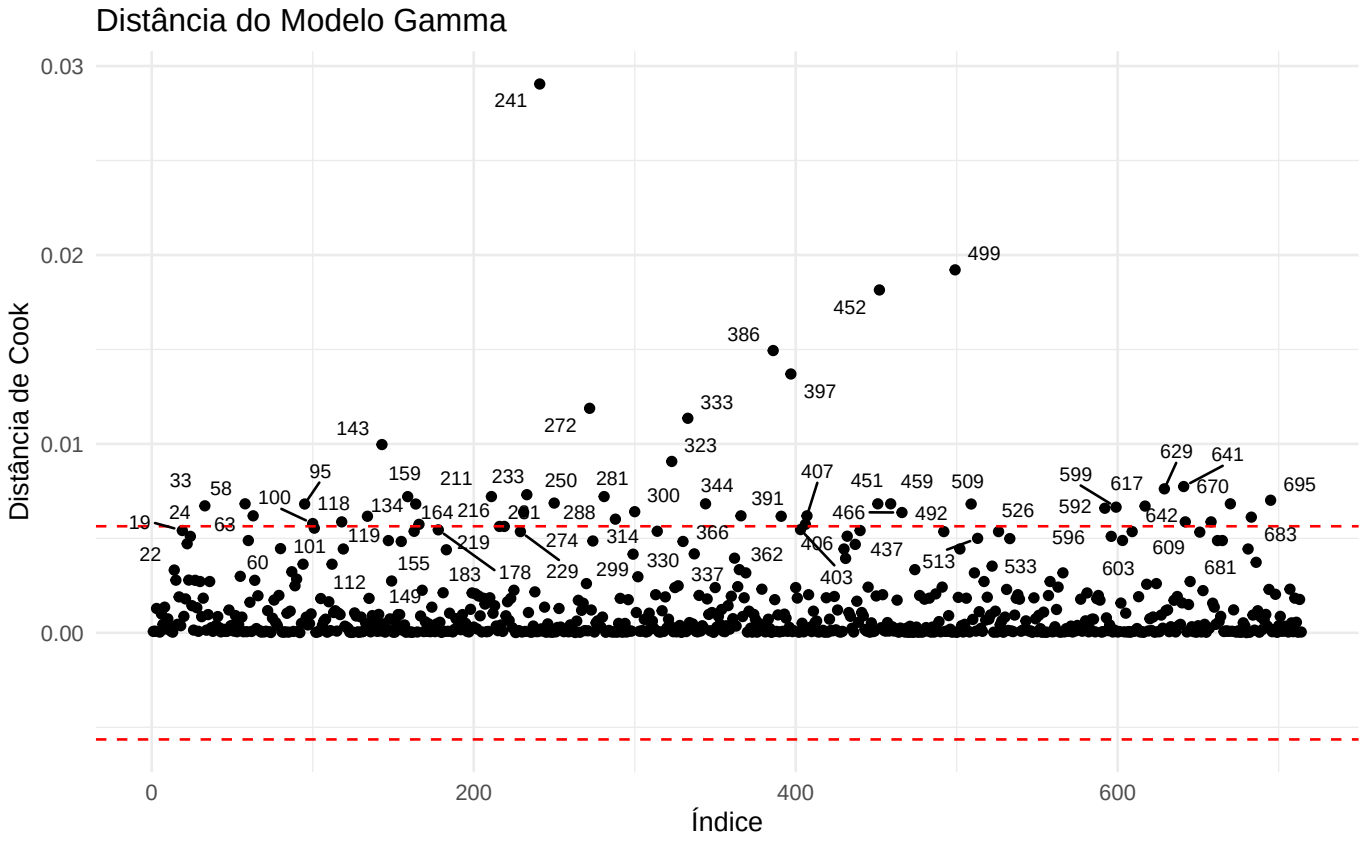

Não observa-se algum resíduo fora dos limites especificados, indicando que não exista pontos de influência.

### 3.2 Envelope Simulado



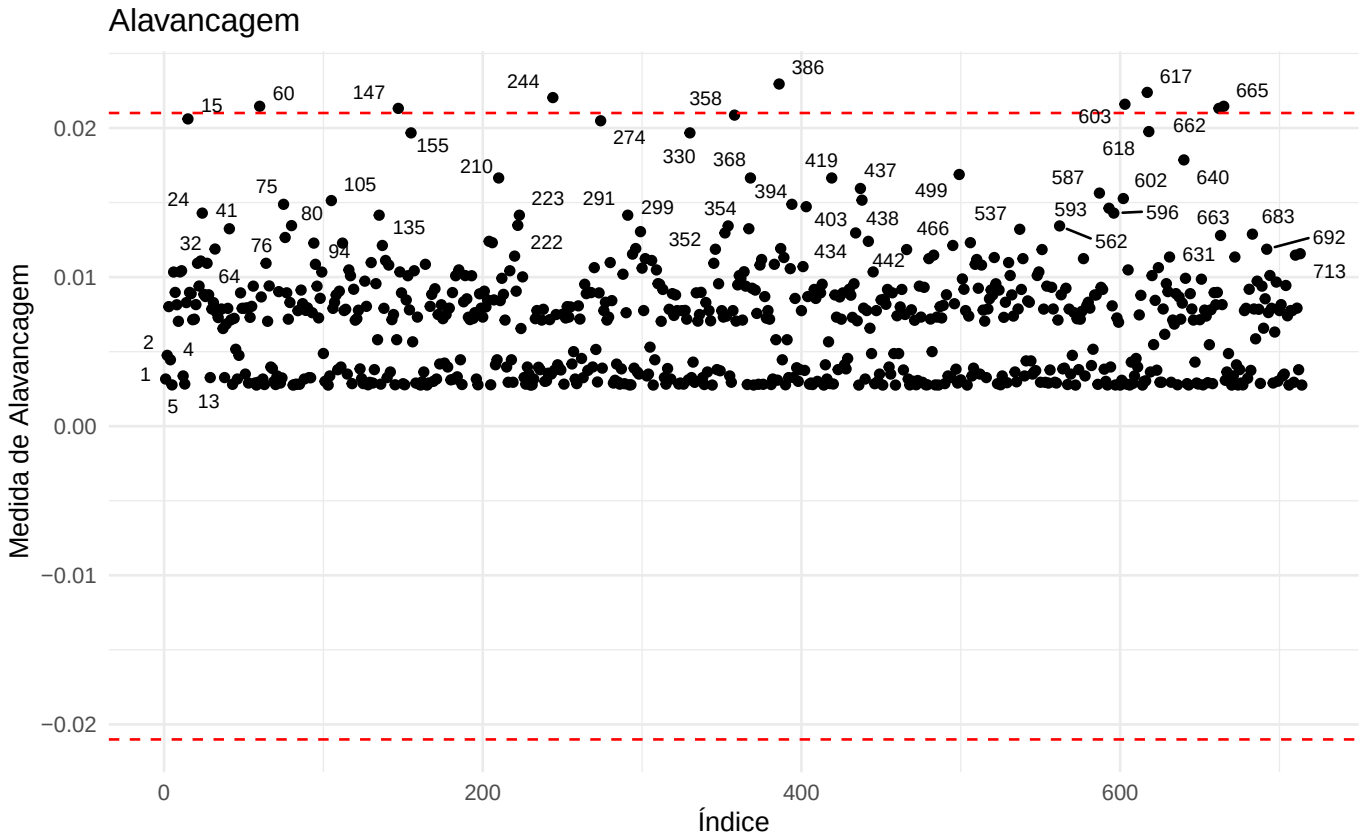

Percebemos que alguns pontos estão fora das bandas simuladas, então devemos procurar por pontos influentes.

### 3.3 Distância de Cook



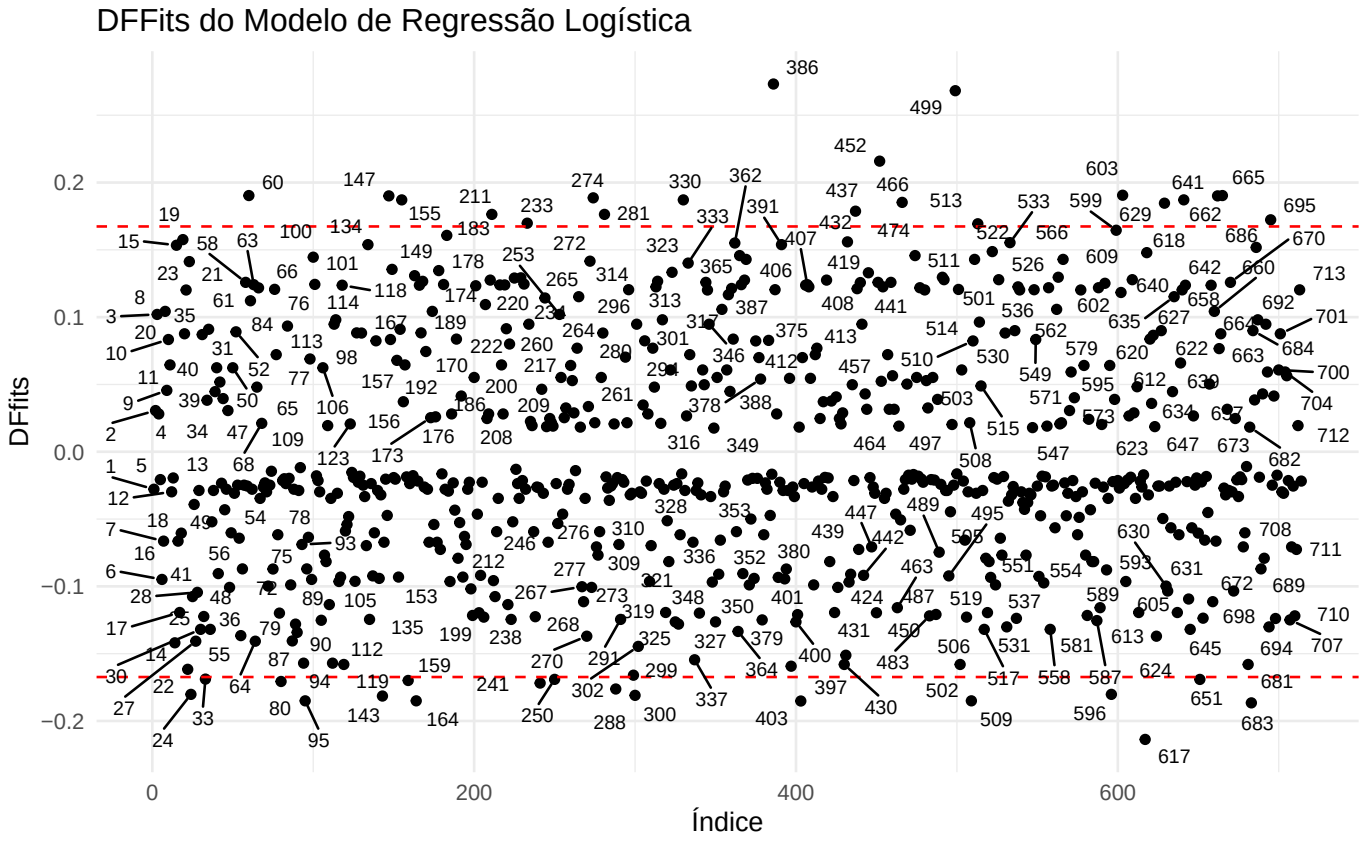

Nota-se que principalmente a observação 241 fica fora dos limites estipulados, com achatamento do gráfico da distância de cook, indicam que são potenciais pontos de influência. Dentre os outros pontos fora das bandas simuladas se destacam as observações 272,333,386,397,452,499.

### 3.4 Alavancagem




Nota-se que alguns pontos ficaram fora dos limites estipulados, sem achatamento do gráfico, mas indicam que são potenciais pontos de influência. Em relação as observações fora dos limites da distância de cook apenas o 386 está fora, destaca-se também o ponto 617.

### 3.5 DFFits




Observamos que os pontos 386 e 499 ficam fora dos limites estipulados de maneira mais contundente com as outras observações ficando próximos aos limites, assim são candidatos a pontos de influência.

## 4 Removendo pontos possivelmente influentes

Nesta seção removeremos alguns pontos possivelmente influentes e faremos as analises para verificar se os pontos candidatos são realmente pontos influentes, verificando como fica a equação do modelo e suas métricas.

### 4.1 Modelo 3

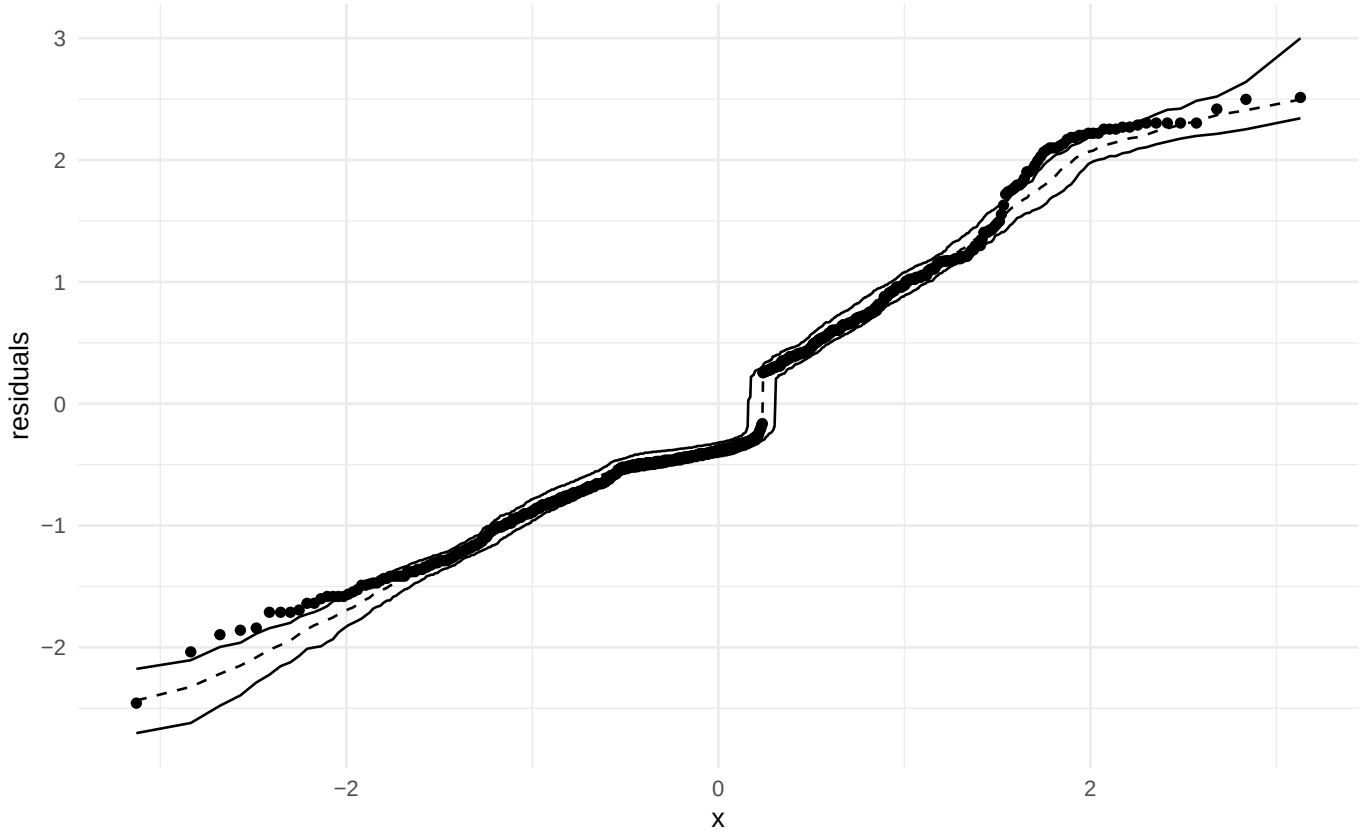
Neste modelo removemos apenas as observações 241 e 386, foi a observação que mais achatou o gráfico e a que mais se repetiu, respectivamente, assim estimaremos os modelos sem elas.

Tabela 9: Métricas de Avaliação do Modelo 3

| .metric | .estimate |
|---|---|
| accuracy | 0.805 |
| kap | 0.594 |
| precision | 0.830 |
| sens | 0.844 |
| spec | 0.747 |

Nas métricas, nota-se uma ligeira melhora em todas.

#### 4.1.1 Envelope Simulado




Porém no Envelope Simulado não nota-se uma diferença expressiva para o correspondente do Modelo 1.

#### 4.1.2 Modelo

Tabela 10: Estatísticas do Modelo 3

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -1.262 | 0.250 | -5.05 | 0 |
| Age | -0.041 | 0.008 | -5.23 | 0 |
| Pclass_X1 | 2.714 | 0.289 | 9.40 | 0 |
| Pclass_X2 | 1.319 | 0.247 | 5.34 | 0 |
| Sex_female | 2.547 | 0.210 | 12.13 | 0 |

Continuamos com todas as covariáveis significativas, tem-se pequenas diferenças nos $\beta$

### 4.2 Modelo 4

Neste modelo, teremos uma aplicação mais radical da análise de influência retirando todos os pontos que contiveram pelo menos um *TRUE* (possível ponto de influência), na função `influence.measures` e posteriomente verificando a equação e as métricas.

Tabela 11: Diferentes medidas de influência

| | dfb.1_ | dfb.Age | dfb.P_X1 | dfb.P_X2 | dfb.Sx_f | dffit | cov.r | cook.d | hat |
|---|---|---|---|---|---|---|---|---|---|
| 58 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 60 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 63 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 101 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 118 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 147 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 163 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 166 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 216 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 219 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 229 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 231 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 241 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 244 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | TRUE |
| 272 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 314 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 323 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 333 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 344 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 358 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 366 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 386 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | TRUE |
| 397 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 406 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 407 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 440 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 451 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 452 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 459 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 492 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 499 | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE |
| 526 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 592 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 603 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 609 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 617 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 642 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 658 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |
| 662 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 665 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE |
| 670 | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE |

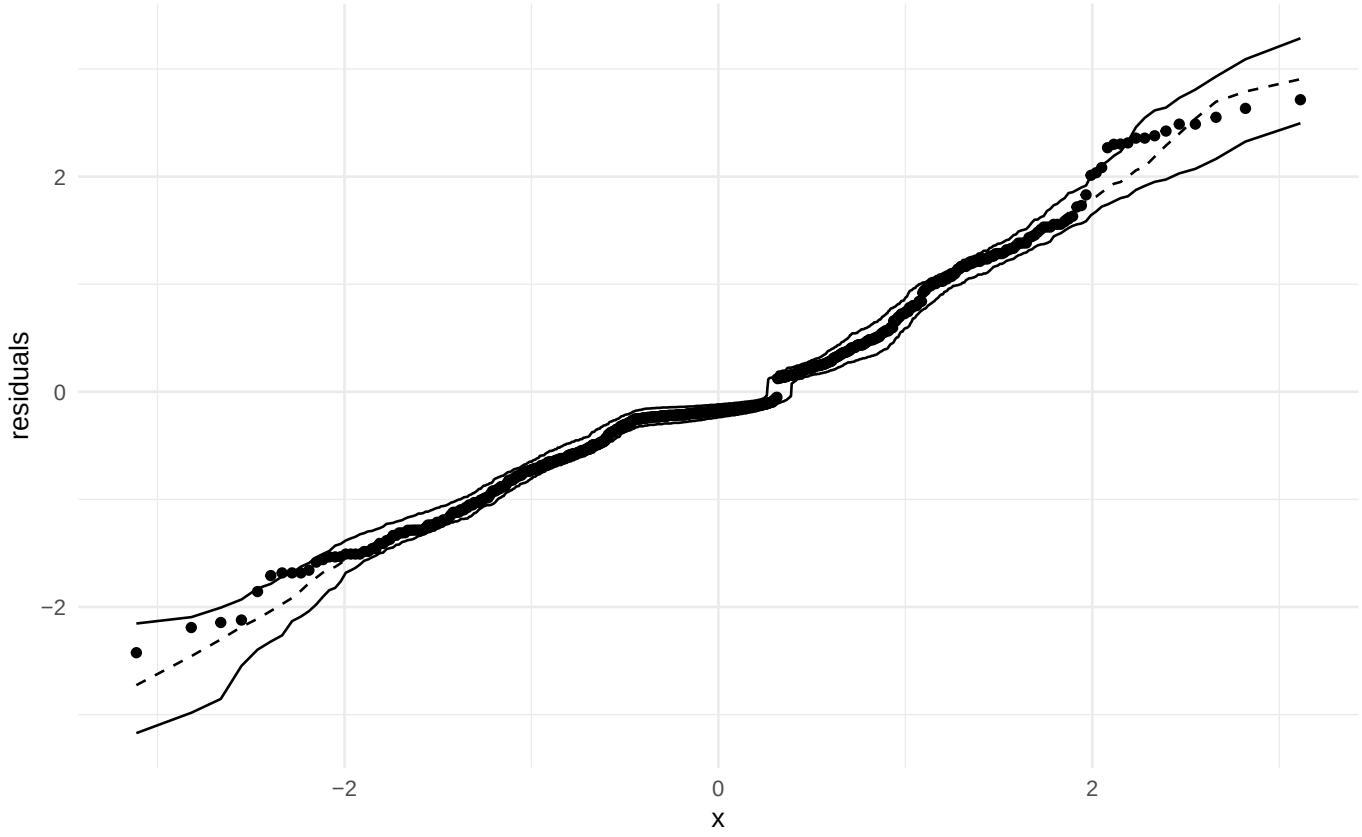
Assim, removeremos todos os pontos acima, pois já estão filtrados, todos que são candidatos a ponto de influência por alguma métrica.

Tabela 12: Métricas de Avaliação do Modelo 4

| .metric | .estimate |
|---|---|
| accuracy | 0.834 |
| kap | 0.643 |
| precision | 0.857 |
| sens | 0.881 |
| spec | 0.756 |

Percebemos uma melhora em relação ao modelo 3, logo temos uma melhora consideravel ao modelo 1.

#### 4.2.1 Envelope Simulado




Reparemos que continua com pontos fora das bandas simuladas, no entanto dimuiu-se a quantidade.

#### 4.2.2 Modelo

Tabela 13: Estatísticas do Modelo 4

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -2.549 | 0.360 | -7.08 | 0 |
| Age | -0.056 | 0.010 | -5.45 | 0 |
| Pclass_X1 | 4.413 | 0.427 | 10.34 | 0 |
| Pclass_X2 | 2.387 | 0.342 | 6.99 | 0 |
| Sex_female | 3.799 | 0.310 | 12.26 | 0 |

Continuamos com todas as covariáveis significativas, porém podemos perceber que cada vez o p-valor é menor e a tem-se diferenças claras nos $\beta's$, aumentos significativos matematicamente em Classe 1, Classe 2 e Sexo Feminino, diminuição no $\beta$ do Intercepto.

# 5 Eu sobreviveria ao Titanic?

Teste em todos os Modelos e testando variando a única classe que não temos certeza a classe, assim nos 4 modelos testei as 3 classes.

*Modelo 1*

Tabela 14: Previsão quanto a sobrevivência no Modelo 1

| .pred | Pclass | Name | Sex | Age |
|---|---|---|---|---|
| 1 | 1 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 2 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 3 | Vítor Bernardo Silveira Pereira | male | 21 |

Então, a previsão para o Modelo 1, eu apenas sobreviveria se fosse na Classe 1.

Tabela 15: Previsão quanto a sobrevivência no Modelo 2

| .pred | Pclass | Name | Sex | Age |
|---|---|---|---|---|
| 1 | 1 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 2 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 3 | Vítor Bernardo Silveira Pereira | male | 21 |

Logo, a previsão para o Modelo 2, eu apenas sobreviveria se fosse na Classe 1, mesma previsão do Modelo anterior.

Tabela 16: Previsão quanto a sobrevivência no Modelo 3

| .pred | Pclass | Name | Sex | Age |
|---|---|---|---|---|
| 1 | 1 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 2 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 3 | Vítor Bernardo Silveira Pereira | male | 21 |

Assim percebemos, a previsão para o Modelo 3, eu apenas sobreviveria se fosse na Classe 1, mesma previsão dos Modelos anteriores.

Tabela 17: Previsão quanto a sobrevivência no Modelo 4

| .pred | Pclass | Name | Sex | Age |
|---|---|---|---|---|
| 1 | 1 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 2 | Vítor Bernardo Silveira Pereira | male | 21 |
| 0 | 3 | Vítor Bernardo Silveira Pereira | male | 21 |

Contudo, concluímos que de acordo com os modelos testados, eu sobreviveria apenas se embarcasse com a classe 1, no entanto seria o mais provável embarcar com a classe 3, de acordo com a situação financeira.

# 6 Razão de Chances

Nesta seção iremos verificar algumas razões de chance muito interessante sobre o desastre do Titanic.

*Modelo Geral*

$$\log\left[\frac{P(\text{Sobreviver})}{1-P(\text{Sobreviver})}\right] = \beta_0 + \beta_1(\text{Idade}) + \beta_2(\text{Classe}_1) + \beta_3(\text{Classe}_2) + \beta_4(\text{Sexo\_feminino})$$

Para encontrar razões de chance, iremos realizar o seguinte procedimento:

$$e^{\log[\frac{P(\text{Sobreviver})}{1-P(\text{Sobreviver})}]} = \frac{P(\text{Sobreviver})}{1-P(\text{Sobreviver})}$$

Primeiramente chegamos em uma chance, para chegar na razão de chance, vamos precisar de outra razão para encontrá-la, logo iremos considerar um modelo, dado que a pessoa que queremos é estimar mulher, e no outro modelo iremos considerar que a pessoa que seja homem:

$$\frac{\frac{P(\text{Sobreviver}|\text{mulher})}{1-P(\text{Sobreviver}|\text{mulher})}}{\frac{P(\text{Sobreviver}|\text{homem})}{1-P(\text{Sobreviver}|\text{homem})}} = \frac{e^{\beta_0+\beta_1(\text{Idade})+\beta_2(\text{Classe}_1)+\beta_3(\text{Classe}_2+\beta_4(\text{Sexo\_feminino})}}{e^{\beta_0+\beta_1(\text{Idade})+\beta_2(\text{Classe}_1)+\beta_3(\text{Classe}_2)}}$$

Aplicando a propriedade da potência que divisão de potencias de bases iguais, pode-se subtrair as potências, e assumindo que todas as outras variáveis são constantes chegamos que a razão de chances é:

$$\text{RC} = e^{\beta_4}$$

Assim, podemos calcular a razão de chance para todos os modelos.

Então temos que as razões de chance para o modelos 1, 2, 3 e 4 são: 12.463, 8.138, 12.771 e 44.641. Então de acordo com os modelos ajustados a chance de pessoas do sexo feminino sobreviveram variam de 8.13 a 44.64 vezes maior do que para pessoas do sexo masculino sobreviverem.

# 7 Conclusão

Observando a equação de todos os modelos, nota-se um padrão, as covariáveis que mais influenciam para sobrevivência são embarcar na primeira classe, ser uma pessoa do sexo feminino e embarcar na segunda classe. Em contrapartida a covariável que menos impacta é a idade, visto que, com maior idade menos provável de sobreviver. Assim, o modelo nos indica que no resgate houve preferência, então, pelas pessoas da primeira classe (mais ricas), mulheres e posteriormente crianças. As mulheres tem incríveis 44 vezes mais chances de sobreviver do que os homens de acordo com o Modelo 4,

já segundo esse modelo as pessoas de primeira classe tem 82.5 vezes mais chances de sobreviver que pessoas em terceira classe. Comparando com a segunda classe, as pessoas que embarcaram na primeira classe tem 7.6 vezes mais chances de sobreviver.

Em relação aos modelos, temos que o modelo que sobressai, com as melhores métricas é o modelo 4, consideravelmente a frente do modelo 1 (inicial) e do modelo 3 (com poucas remoções de pontos influentes), percebemos que o modelo 2 é o que se sai pior nessas métricas, também podemos avaliar os critérios de seleção do modelo:

Tabela 18: Critérios de Seleção do Modelo para as Regressões Logisticas

| null.deviance | df.null | logLik | AIC | BIC | deviance | df.residual | nobs |
|---|---|---|---|---|---|---|---|
| 965 | 713 | -324 | 657 | 680 | 647 | 709 | 714 |
| 962 | 711 | -318 | 647 | 670 | 637 | 707 | 712 |
| 894 | 674 | -222 | 453 | 476 | 443 | 670 | 675 |

Assim, percebemos que nos principais critérios de seleção AIC e BIC, o modelo 4, tem medidas consideravelmente menores e melhores. Para a Regressão Logística - `glmnet`, temos:

Tabela 19: Critérios de Seleção do Modelo - GLMNET

| nulldev | npasses | nobs |
|---|---|---|
| 965 | 398 | 714 |

Assim comparando esses critérios com o da Tabela acima, podemos perceber que os outros modelos são superiores. Então para a seleção do melhor modelo temos que levar em consideração alguns fatores, que o Modelo 4 foi o melhor modelo no geral, em métricas de acurácia e critérios de seleção, logo o melhor modelo para previsão dos dados.

No entanto, o Modelo 1, com todos os dados pode ser o mais correto inferencialmente, pois seria o mais representativo em relação a população, também pode estar sendo mais cauteloso quanto ao pressuposto de amostra aleatória, o que no Modelo 4 quebramos com a remoção das observações.

Tendo essas observações em vista e sem um estudo mais aprofundado para a remoção de pontos influentes, a escolha para o Modelo 4, seria para um modelo com melhor capacidade preditiva e a escolha para o Modelo 1, seria para um modelo com melhores conclusões inferenciais.