Um Passeio nas Avaliações de Animes

Alisson Rosa

Resumo

O site denominado MyAnimeList, é uma rede social focada nos consumidores de animes e mangás, na qual possui como maior característica a possibilidade de seus usuários criarem uma lista pessoal para que possam catalogar as obras e classificar-las através de notas, dessa maneira o presente trabalho apresenta uma análise descritiva dos animes como também é desenvolvido um modelo de regressão para modelar o score que os usuários outorgam ao anime.

Sumário

1 Uma Introdução aos Animes

Você sabe o que são os animes? Com certeza você já deve ter ouvido falar nesse termo. Acontece que eles nada mais são do que desenhos animados produzidos no Japão. A palavra em si é uma abreviação de "animação" em japonês. Anime nunca esteve na mesma caixinha do desenho. A estrutura da narrativa é diferente. Enquanto desenhos em geral seguem a cada episódio uma história, no anime você já tem mais um aspecto de novela, em que exige o envolvimento. Você precisa esperar pelo próximo capítulo para ver o desenrolar do que aconteceu no último. Tem o acompanhamento. Não é acompanhar só por gostar, é acompanhar a narrativa.

Se você abrir hoje a Netflix, vai demorar um tempo até encontrar algo que você queira ver, mas vai achar. Você pode achar algo que te interesse pela narrativa, ou algo totalmente fora da curva. Os animes sempre foram assim: um grande catálogo, que navegando você encontra narrativas interessantes e aquelas que fazem sua cabeça explodir. Existe a base, onde estão grande parte dos animes. E existem os atemporais. É como no cinema. Não tem aquela massiva leva de filmes que estão aí e nunca clicamos? E não tem aquelas recomendações certeiras? Existem animes que souberam explorar recursos metalinguísticos de seu formato, para contar uma história ainda maior.

Dessa maneira surge o MyAnimeList, muitas vezes abreviado para MAL, que é uma rede social focada nos consumidores de animes e mangás, na qual possui como maior característica a possibilidade de seus usuários criarem uma lista pessoal para que possam catalogar as obras e classificar-las através de notas. À vista disso vamos analisar os *reviews* deixados pelos usuários até o ano de 2020, a base aqui utilizada e modificada pode ser encontrada clicando-se aqui.

2 Estimando Quantidades Básicas

Como dito anteriormente, existe uma quantidade vasta de animes disponíveis, assim a análise será reduzida a aqueles que possuem ao menos 15.000 membros no MAL. Nessa seção, vamos responder algumas perguntas básicas, veremos o comportamento da quantidade de episódios e membros.

2.1 Episódios

É sempre uma informação relevante a quantidade de episódios de um anime, pois por exemplo, existem alguns com mais de 1000 episódios[1] levando assim um bom tempo para serem finalizados, dessa maneira vamos avaliar o comportamento dos episódios de forma geral.

Tabela 1: Comportamento da Quantidade de Episódios

variable	mean	median	sd	min	max	na_count
episodes	14.3	12	24.3	1	500	0

Portanto a Tabela 1 informa que depois do banco ser ajustado, o anime com a maior quantidade de episódios é Naruto: Shippuuden, nota-se pela Tabela 2 também algo que era de se esperar, uma assimetria na quantidade de episódios, onde a maioria dos animes tem quantidade de epis[odios] totais em torno de 12, 13, 24 e 26, um fato esperado pois as temporadas normalmente são divididas em duas, geralmente possuindo 12 ou 13 episódios cada parte.

Tabela 2: Top 4 Quantidade de Episódios

episodes	n
12	881
13	372
26	189
24	165

2.2 Membros

Existem animes que são uma explosão de fãs, e uma boa maneira de avaliar isso é vendo a quantidade de membros, uma visão geral dos membros pode ser visto pela Figura 1

[1] O famigerado One Piece é um ótimo exemplo

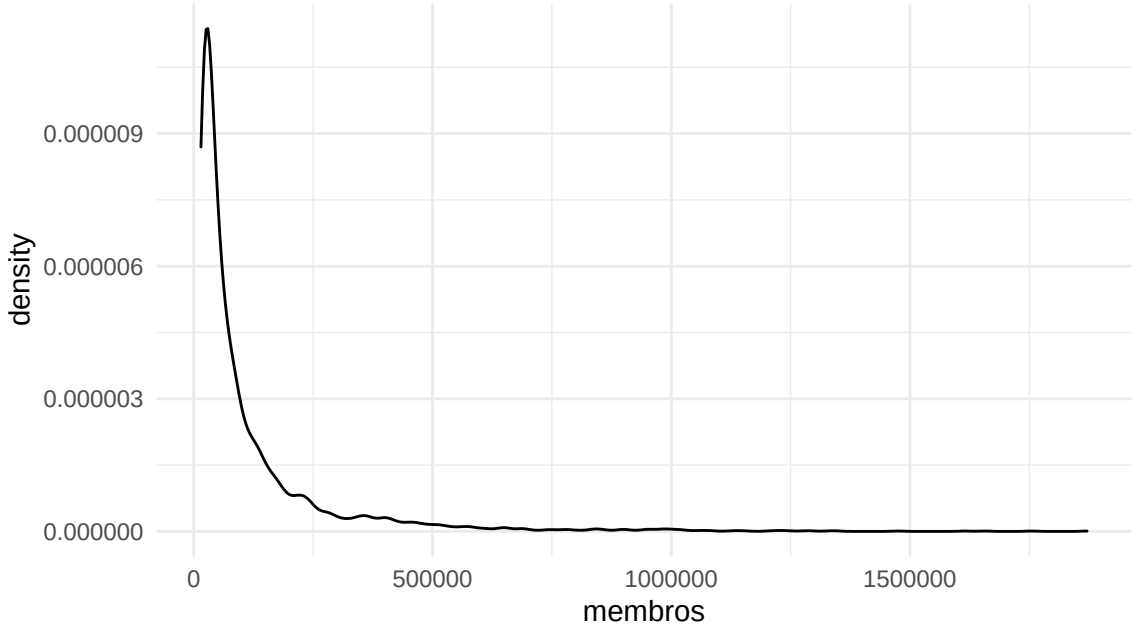

Figura 1: Densidade da Quantidade de Membros

O que podemos ver é claramente uma assimetria, algo também esperado, pois são poucos animes que tem a capacidade de fazer um grande sucesso. Vejamos o top 3 em quantidade de membros.

Tabela 3: Top 3 animes em quantidade de Membros

title	episodes	members	score
Death Note	37	1871043	0.865
Shingeki no Kyojin	25	1754979	0.847
Sword Art Online	25	1657823	0.749

O primeiro é o famoso Death Note[2] um dos animes que é categorizado como "porta entrada" para esse mundo, pois trata de temas importantes que permeiam a humanidade a um bom tempo. O segundo lugar fica para Shingeki no Kyojin (SNK) [3], uma obra que para muitos é considerada uma das melhores histórias dos últimos anos, pois não segue o padrão de "storytelling" dos animes atuais[4]. E por último esta Sword Art Online, um anime que possui uma grande quantidade de fãs e *haters*, mas por conter elementos de *games* possui uma grande quantidade de fãs desse género.

2.3 Top 50 Popularidade

Todo ano existem animes que se destacam, sejam lançamentos ou sejam temporadas novas, dessa maneira é interessante verificar a influência da popularidade no *score*, o que podemos ver pela Figura 2

[2]O preferido pelo autor que vos escreve
[3]Attack on Titan em inglês
[4]A última temporada de SNK está marcada para ser lançada em 2023!!

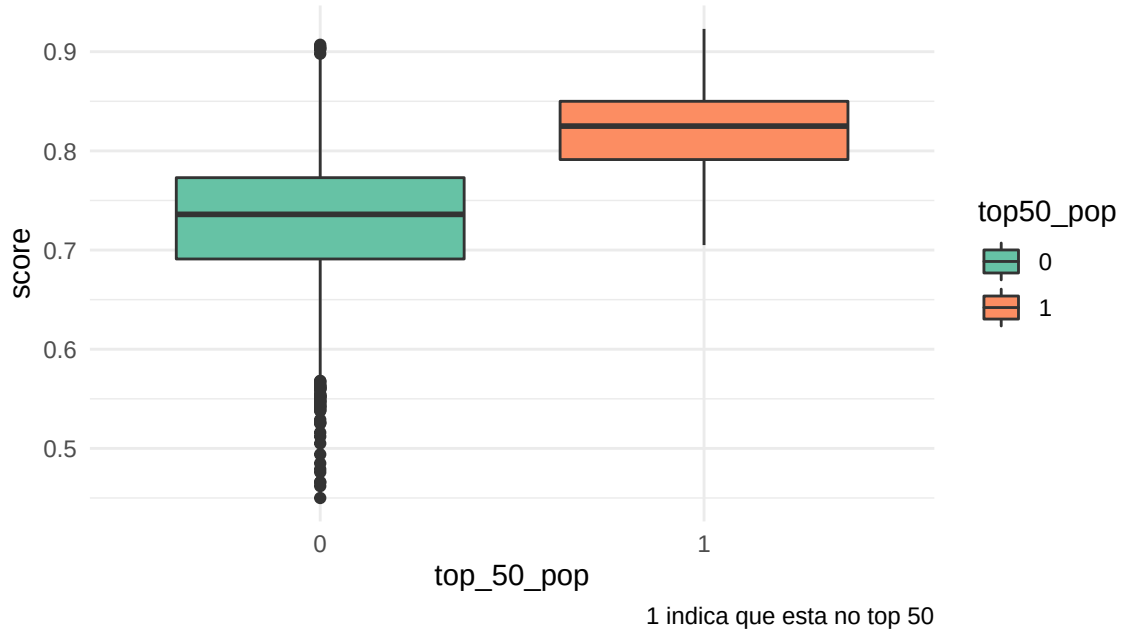

Figura 2: Comparação dos top 50 em popularidade em referência ao score

Podemos notar que aqueles que se destacam para a comunidade em termos gerais possuem *score* maior, um fato importante para a modelagem do score em seções seguintes.

3 O Modelo de Regressão Beta

Seja Y uma variável aleatória seguindo a distribuição beta, é possível reparametrizarmos a distribuição em termo de sua média (μ), dessa maneira se construirmos um modelo de regressão para a média, estaremos construindo um modelo para um parâmetro da distribuição, é nesse sentido que entra em cena o modelo de regressão beta, definido em [1].

Sejam $Y_1, \dots Y_n$ variáveis aleatórias que seguem a distribuição beta reparametrizada definida em [1] com cada Y_t possuindo uma média μ_t, o modelo assume que cada μ_t pode ser escrito da seguinte maneira:

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i \tag{1}$$

em que os $\beta's$ são parâmetros a serem estimados, os x_s são variáveis observadas e g(.) é uma função estritamente monótona e duas vezes diferenciável.

Dessa maneira a variável de interesse (Y_t) aqui é o *score*, e as covariáveis (x_t) são:

- Quantidade de Episódios (episodes)
- Quantidade de Membros (members)
- Quantidade de episódios categorizada (eps):
  - Menos de 12 - Very low
  - Entre e 12 e 352 - Ok
  - Do contrário - OP[5]
- Quantidade de Membros categorizada (membros):
  - Menos do que 100000 - low

[5]One Piece!

  - Do contrário - insane

- Top 50 em termos de Popularidade (top_50_pop)
  - Variável que informa se o anime está entre os top 50 de popularidade.

Ajustando o modelo tomando g = logit[6], temos como resumo a Tabela 4

Tabela 4: Coeficientes Estimados

	Estimate	Std. Error	P.value
(Intercept)	1.048	0.020	<0.001
episodes	0.001	0.000	<0.001
members	0.000	0.000	<0.001
epsOP	-0.214	0.013	<0.001
epsvery-low	0.065	0.013	<0.001
membroslow	-0.151	0.016	<0.001
top50_pop	-0.210	0.055	<0.001

Nota-se que todas variáveis são significativas para predição de score a qualquer nível de significância razoável, os coeficientes associados a animes com muitos episódios e com poucos membros são negativos, indicando que tendem a baixar o score,

Porém o coeficiente associado aos top 50 animes em popularidade também é negativo. Um fato que não corresponde ao que era esperado pela Figura 2, porém na Figura 2 avalia-se a variável isoladamente, no modelo para interpretarmos o coeficiente de uma variável fixamos os valores das outras, portanto a correlação existente entre as variáveis influência no valor dos coeficientes estimados.

3.1 Medidas de Adequação

O modelo de regressão de beta não se adequa a qualquer situação, assim como qualquer modelo estatístico clássico devemos fazer uma avaliação geral da adequação. Dessa maneira vamos avaliar os gráficos dos resíduos, indices e valores ajustados.

O gráfico de resíduo e indices deve ter um comportamento aleatório e possuir poucos valores acima de -3 e 3, para o modelo ajustado tal fato pode ser verificado pela Figura 3

[6] $\text{logit(x)} = \log(\frac{x}{1-x})$

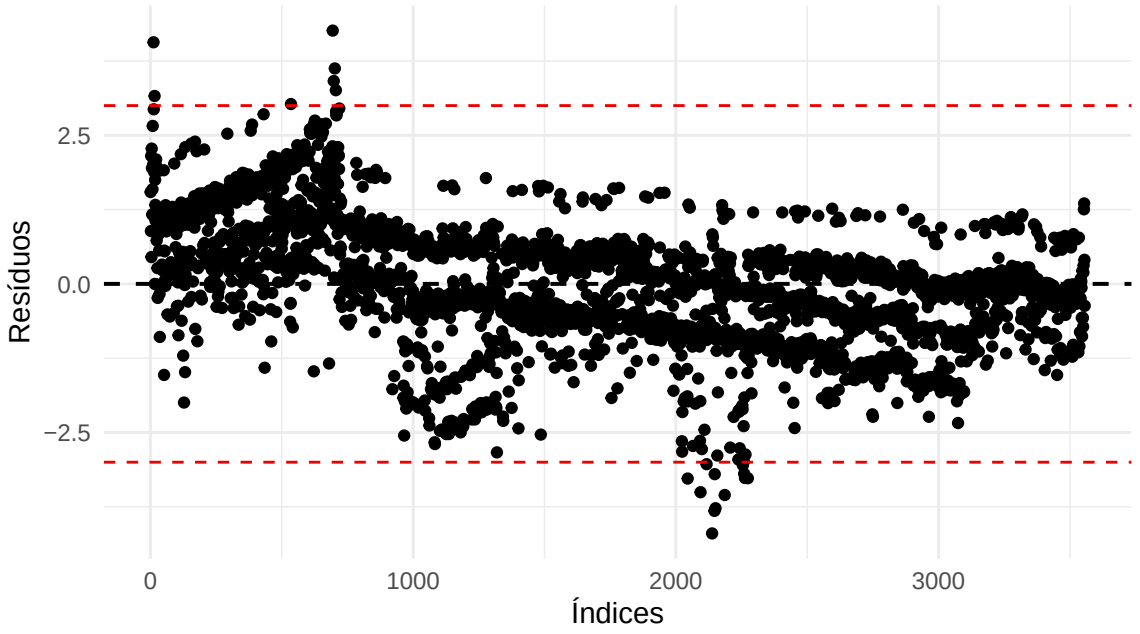

Figura 3: Indices vs Resíduos

O gráfico de medidas de alavancagem e da distância de cook em geral não podem ter pontos extravagantes em termos de valores observados, o que pode ser visto pela Figura 4 é que existem duas observações claramente se destacando, porém são os animes Naruto Shippuden[7] e Sword art Online, dessa maneira esses animes não serão removidos da modelagem.

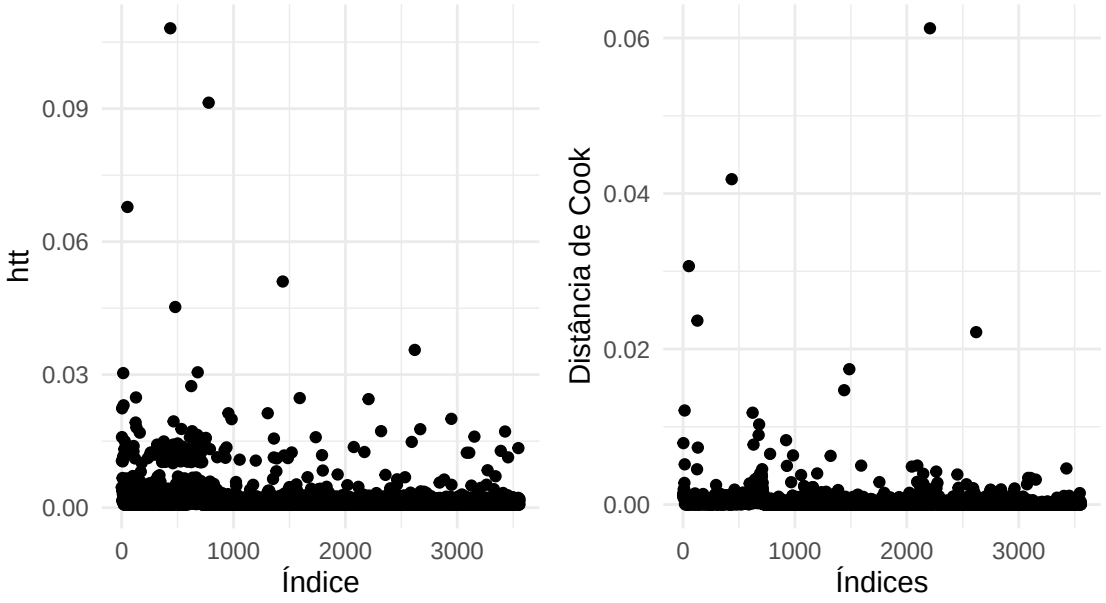

Figura 4: index vs hatvalues

[7]Um verdadeiro clássico

4 As Evidências e As Conclusões

Vimos nesse breve ensaio que a quantidade de episódios de anime flutua entre 12 e 25, a distribuição de membros é plenamente assimétrica, dessa maneira o que podemos concluir é que existem inúmeros animes existentes, porém poucos conseguem se destacar, e esses que se destacam em termos de modelagem a análise de diagnóstico tende a destacar como pontos aberrantes, o que foi visto pela Figura 4, isso evidência que não se deve remover observações de banco de dados sem antes saber precisamente quão informação elas carregam.

5 Código

```
library(tidyverse)
library(fastrep)
library(betareg)
library(patchwork)
df <- read.csv("data/animes.csv")


df <- df |>
  drop_na() |>
  distinct() |>
  filter(
    title != "Doraemon (1979)",
    title != "Tenkuu Danzai Skelter+Heaven",
    title != "Hametsu no Mars",
    title != "Pupa"
  ) |>
  select(-synopsis, -uid, -genre, -aired, -img_url, -link) |>
  mutate(score = score / 10) |>
  mutate(
    eps = case_when(
      episodes < 12 ~ "very-low",
      (12 < episodes & episodes < 352) ~ "ok",
      TRUE ~ "OP"
    ),
    membros = case_when(
      members < 100000 ~ "low",
      TRUE ~ "insane"
    ),
    top50_pop = ifelse(popularity < 100, 1, 0)
  ) |>
  filter(members > 15000)
df |>
  select(episodes) |>
  describe() |>
  tbl("Comportamento da Quantidade de Episódios")
df |>
  tibble() |>
  group_by(episodes) |>
  summarise(n = n()) |>
  arrange(desc(n)) |>
  filter(episodes != 1) |>
  head(4) |>
  tbl("Top 4 Quantidade de Episódios")
```

```r
df |>
  ggplot(aes(x = members)) +
  geom_density() +
  labs(x = "membros", y = "density")
df |>
  arrange(desc(members)) |>
  head(3) |>
  select(-popularity, -ranked, -eps, -membros, -top50_pop) |>
  tbl("Top 3 animes em quantidade de Membros")



df |>
  mutate(top50_pop = factor(top50_pop)) |>
  ggplot(aes(x = top50_pop, y = score, fill = top50_pop)) +
  geom_boxplot() +
  labs(x = "top_50_pop", caption = "1 indica que esta no top 50")
fit <- betareg(score ~ .,
  data = df |> select(-title, -ranked, -popularity), x = TRUE, link = "logit"
)
teste <- fit |> summary()
teste$coefficients$mean |>
  as.data.frame() |>
  mutate(P.value = format.pval(`Pr(>|z|)`, eps = 0.001)) |>
  select(-`Pr(>|z|)`, -`z value`) |>
  tbl("Coeficientes Estimados")
residuot2 <- residuals(fit, type = "sweighted2")

yajust <- fitted.values(fit)
yhat <- hatvalues(fit)
dcook <- cooks.distance(fit)

deviance <- sum(residuals(fit, tipe = "deviance")^2)
df <- df |> bind_cols(index = 1:nrow(df))
df |> ggplot(aes(
  x = index,
  y = residuot2
)) +
  geom_point(size = 1.5) +
  geom_hline(
    yintercept = 3, colour = "red2",
    size = 0.5, linetype = "dashed"
  ) +
  geom_hline(
    yintercept = 0, colour = "black",
    size = 0.7, linetype = "dashed"
  ) +
  geom_hline(
    yintercept = -3, colour = "red2",
    size = 0.5, linetype = "dashed"
  ) +
  labs(x = "Índices", y = "Resíduos")
ggplot(df, aes(x = index, y = yhat)) +
```

```
  geom_point(size = 1.5) +
  labs(x = "Índice", y = "htt") +
  ggplot(df, aes(
    x = index,
    y = dcook
  )) +
  geom_point(size = 1.5) +
  labs(x = "Índices", y = "Distância de Cook")
```

6 Bibliografia

[1] Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. Journal of applied statistics. 2004;31:799–815.

[2] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Available from: https://www.R-project.org/.

[3] Rosa A. Fastrep: Time-saving package for creating reports. 2022.