

UNIVERSIDADE FEDERAL DO PARANÁ

André Luiz Grion – GRR20159284

Bruno Henrique Abreu – GRR20159983

Maria Tereza Neves de Oliveira – GRR20159323

**DETERMINAÇÃO DA RECEITA BRUTA DOS FILMES  
NORTE-AMERICANOS**

CURITIBA

Agosto de 2017

## Resumo

O presente trabalho tem como objetivo identificar as características mais influentes na determinação da receita bruta dos filmes norte-americanos nos últimos anos. O banco de dados foi retirado da plataforma Kaggle, contendo originalmente 5043 observações (filmes) e 28 variáveis, extraídas do website IMDB. Foram aplicados diversos filtros no intuito de homogeneizar as informações e através de técnicas estatísticas como regressão linear múltipla e transformações de variáveis, fez-se a modelagem dos dados. Ao final, foi possível escolher as variáveis de maior contribuição e encontrou-se um modelo que explicou aproximadamente 70% da variação da receita bruta de filmes norte-americanos.

**Palavras-chave:** 1. Receita Bruta 2. Filmes 3. IMDB 4. Regressão Linear

# Conteúdo

**Resumo**

**Introdução** 3

**Material e Métodos** 3

Conjunto de dados . . . . .	3
Regressão linear múltipla . . . . .	5

**Resultados e Discussão** 6

**Conclusão** 11

**Anexo** 12

# Introdução

IMDb, sigla que significa Base de Dados de Filmes da Internet (em Inglês: *Internet Movie Database*), é um dos mais respeitados sites de crítica popular de Filmes e Séries do mundo. O site conta com mais de 4,4 milhões de títulos, cada um com centenas de atributos, divididos em categorias.

Desse enorme banco de dados, através de um pacote do programa Python, foram extraídas 28 variáveis de 5043 observações (filmes). Esses dados foram disponibilizados na plataforma Kaggle em 30 de agosto de 2016, portanto, os dados analisados se referem no máximo a essa data.

A variável resposta escolhida no presente trabalho foi receita bruta (**gross**) e como variáveis explicativas escolheram-se orçamento, duração, gênero, diretor, entre outras. No intuito de otimizar o estudo, alguns filtros foram aplicados à base de dados, bem como transformações de escala de variáveis, métodos de seleção apropriados e diagnósticos de influência para o correto ajuste do modelo.

# Material e Métodos

## Conjunto de dados

Os dados utilizados no modelo de regressão linear foram retirados da plataforma Kaggle, <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset/version/1>, disponibilizados em 30 de agosto de 2016, sendo portanto dessa data, o último registro de dados. A princípio, o banco de dados continha 5043 observações (filmes) e 28 variáveis. As variáveis contidas no modelo estão descritas na Tabela 1.

Tabela 1: Descrição das variáveis presentes no banco de dados

#	Variáveis	Label
1	color	Filme colorido ou preto e branco
2	director_name	Diretor do filme
3	num_critic_for_reviews	Número de críticas por avaliação
4	duration	Duração do filme
5	director_facebook_likes	Quantidade de Likes do diretor do filme no facebook
6	actor_3_facebook_likes	Quantidade de Likes do ator 3 do filme no facebook
7	actor_2_name	Nome do ator 2 do filme
8	actor_1_facebook_likes	Quantidade de Likes do ator 1 do filme no facebook
9	gross	Renda bruta do filme
10	genres	Genêro do filme
11	actor_1_name	Nome do ator 1 do filme
12	movie_title	Nome do Filme
13	num_voted_users	Número de votos dos usuários do IMDB
14	cast_total_facebook_likes	Total de Likes do elenco no Facebook
15	actor_3_name	Nome do ator 3 do filme
16	facenumber_in_poster	Números de faces no pôster
17	plot_keywords	Palavras-chave do enredo

#	Variáveis	Label
18	movie_imdb_link	Link do filme no IMDB
19	num_user_for_reviews	Número de usuários por avaliação
20	language	Língua do filme
21	country	País do Filme
22	content_rating	Classificação do Filme
23	budget	Orçamento do filme
24	title_year	Ano que o filme foi lançado
25	actor_2_facebook_likes	Quantidade de Likes do ator 2 do filme no facebook
26	imdb_score	Nota no IMDB
27	aspect_ratio	Proporção da tela
28	movie_facebook_likes	Quantidade de Likes do filme no facebook

O banco de dados possui filmes de mais de 60 países diferentes de diferentes idiomas. Percebe-se também que a quantidade de filmes se estabiliza à partir de 1999 (Figura 1). Por essas razões definiu-se a população alvo: filmes a partir de 1999, dos EUA e em língua inglesa.

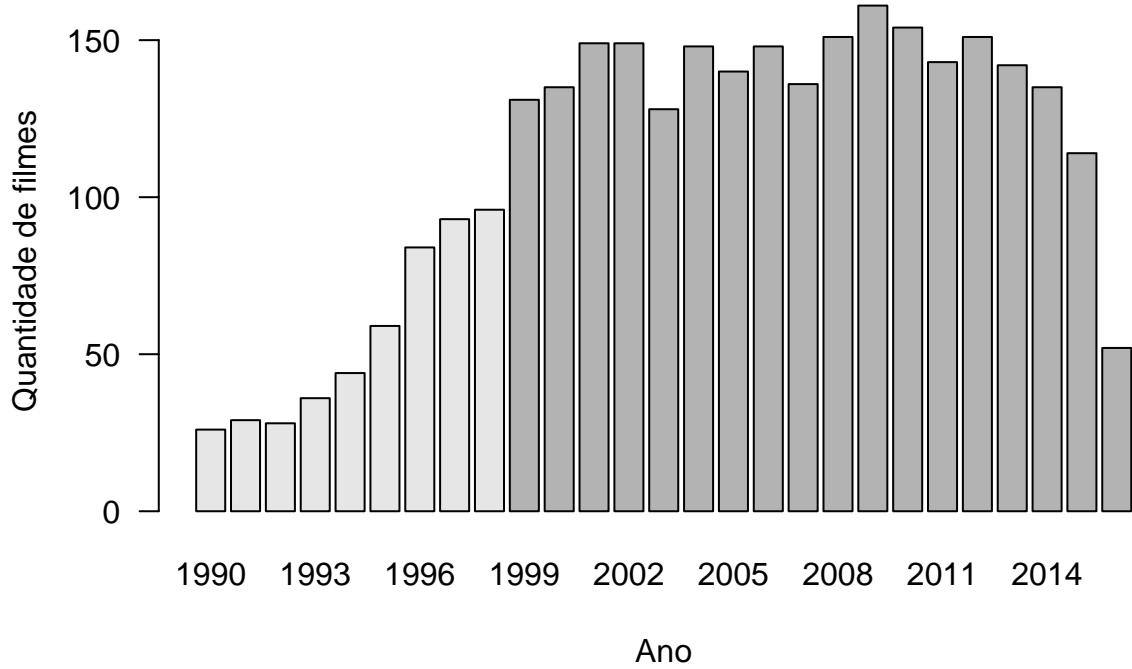


Figura 1: Volumetria de filmes a partir de 1990

Filtrando os filmes a partir de 1999, restam  $n = 2467$  filmes, ou seja, 48,92% da base de dados original.

Variáveis relacionadas com o facebook foram descartadas pelos problemas descritos no próprio repositório do banco de dados. Segundo o mantenedor, zero *likes* não significa, necessariamente, que a página do diretor, atores e/ou filme não receberam nenhum *like*. Pode ser somente problema no algoritmo que fez essa busca.

Também foram encontradas divergências nas variáveis `actor_*_name`, pois os critérios para

escolha dos atores de cada filme não foram explicitadas pelo mantenedor do banco de dados. Essas variáveis foram, portanto, eliminadas do banco de dados.

Embora os filmes do banco de dados apresente somente 22 gêneros diferentes, alguns filmes são classificados em subgêneros (ver Reindeer Games). Uma classificação Action|Adventure pode apresentar um efeito diferenciado de um filme classificado somente como Action, por exemplo, por esse motivo optou-se por manter a classificação original. Devido a grande variedade de combinações, no entanto, considerou-se os efeitos das combinações de gêneros com frequência de ao menos 15 filmes. Gêneros e combinações de gêneros com frequência menor que 15 foram todos agrupados na categoria **Others**.

O efeito individual de diretor foi testado somente para 34 diretores que dirigiram em média mais de 1 filme a cada 3 anos no período, ou seja, dirigiram 7 ou mais filmes no período analisado. Diretores altamente requisitados podem ser um indício de que os mesmos tragam maior retorno financeiro. Os outros 2365 diretores presentes na base de dados que dirigiram, cada um, menos de 7 filmes foram agrupados na categoria **Others**. Desses 2365 diretores que não tiveram o efeito individual estimado, 89,18% dirigiram somente 1 ou 2 filmes no período, o que inviabilizaria a estimativa individual, justificando o agrupamento.

Na Tabela 2 é apresentada uma estatística descritiva das variáveis numéricas consideradas para análise.

Tabela 2: Estatística descritiva das variáveis numéricas consideradas para análise

	Média	D.P.	Mediana	Min	Max	N
duration	107,30	19,16	104,0	46,000	280	2466
imdb_score	6,30	1,03	6,4	1,600	9	2467
num_voted_users2	103,41	148,53	51,9	0,022	1676	2467
num_user_for_reviews	340,66	422,07	207,0	1,000	4667	2466
num_critic_for_reviews	182,98	129,31	155,0	1,000	813	2464
facenumber_in_poster	1,48	2,25	1,0	0,000	43	2461
budget2	44,34	47,22	29,5	0,000	300	2324
gross2	55,61	73,73	31,9	0,001	761	2467

## Regressão linear múltipla

Foi ajustado um modelo de regressão linear múltipla:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim NM_n(\mathbf{0}; \sigma^2 \mathbf{I}) \end{aligned}$$

onde  $\mathbf{y}$  é o vetor da variável resposta de interesse (**gross**),  $\mathbf{X}$  é a matrix do modelo com  $n$  linhas e números de variáveis explanatórias + 1 colunas,  $\boldsymbol{\beta}$  é o vetor 1 + o número de parâmetros associados cada variável explanatória,  $\boldsymbol{\epsilon}$  é o vetor dos  $n$  erros aleatórios associados a cada observação,  $\mathbf{0}$  é o vetor nulo de dimensão  $n$ ,  $\mathbf{I}$  é a matrix identidade  $n \times n$  e  $NM_m$  denota a distribuição normal multivariada de dimensão  $n$ .

## Resultados e Discussão

Na Figura 2 são apresentados os gráficos de dispersão (fora da diagonal) e densidade empírica (diagonal) das variáveis numéricas presentes no banco de dados.

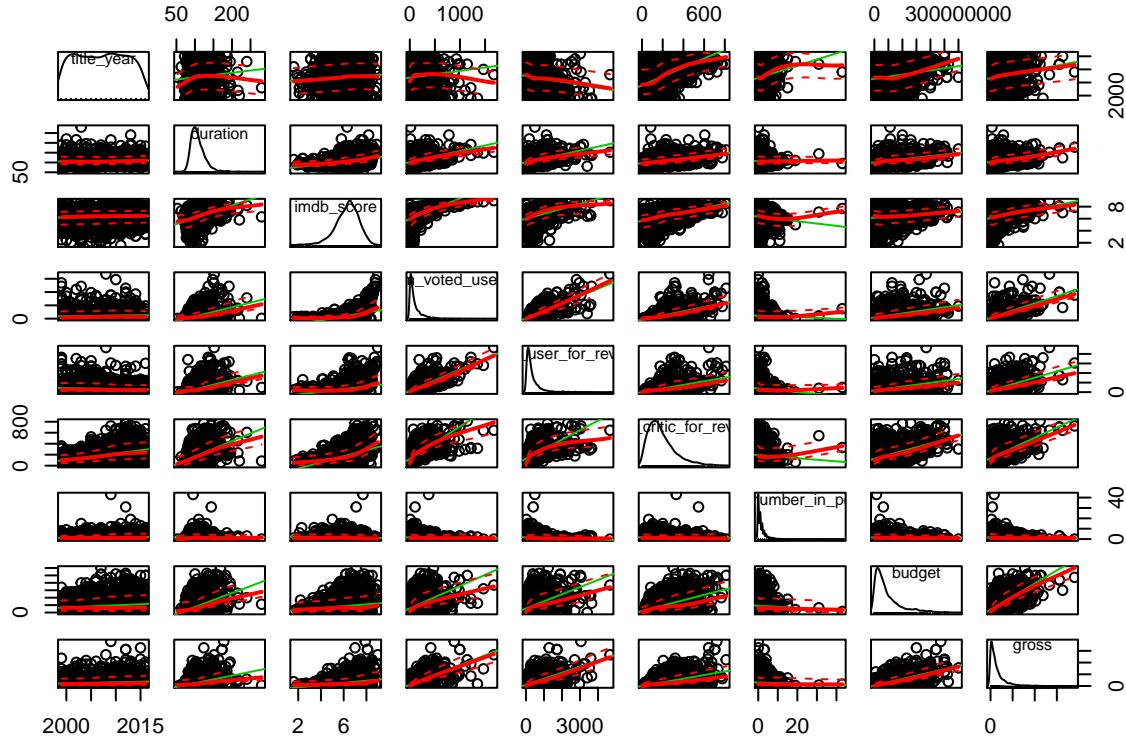


Figura 2: Densidade empírica (diagonal) e associação entre variáveis numéricicas consideradas no modelo

Os gráficos de colunas com as frequências dos níveis das variáveis categóricas consideradas para análise são apresentados na Figura 3.

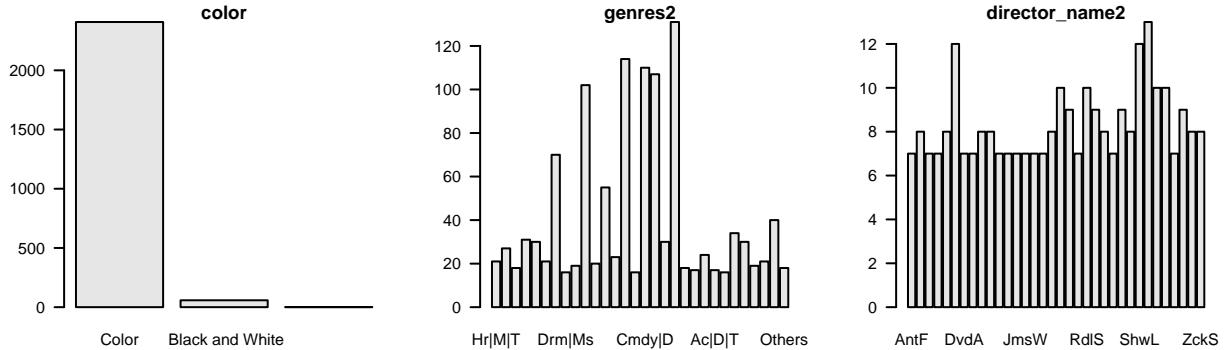


Figura 3: Distribuição das variáveis categóricas consideradas no modelo e associação com variável receita bruta

Para composição do modelo, foram utilizadas os métodos de seleção *backward*, *forward* e *stepwise*. Todos os métodos selecionaram as mesmas variáveis. Partindo do modelo completo, as variáveis retiradas (com menores contribuições utilizando critério de informação de Akaike) foram:

- `imdb_score`
- `facenumber_in_poster`
- `color`

Partindo do modelo somente com intercepto, as variáveis incluídas (com maiores importâncias), foram, respectivamente:

- `budget2`
- `num_voted_users2`
- `director_name2`
- `genres2`
- `num_user_for_reviews`
- `duration`
- `title_year`
- `num_critic_for_reviews`

Na Tabela 3 de análise de variância, utilizando soma de quadrados do tipo II, se confirma o exposto pelos métodos de seleção de variáveis utilizados. As variáveis `imdb_score`, `facenumber_in_poster` e `color` não contribuem para melhoria da explicação da variável resposta na presença das outras variáveis.

Tabela 3: Análise de variância com soma de quadrado tipo II  
do modelo com variáveis explanatórias selecionadas

	Sum Sq	Df	F value	Pr(>F)
color	3228	1	1,759	0,185
genres2	152911	30	2,777	0,000
director_name2	307355	34	4,924	0,000
title_year	12811	1	6,979	0,008
duration	9192	1	5,007	0,025
imdb_score	58	1	0,032	0,859
num_voted_users2	553432	1	301,484	0,000
num_user_for_reviews	46941	1	25,571	0,000
num_critic_for_reviews	5207	1	2,836	0,092
facenumber_in_poster	231	1	0,126	0,723
budget2	1383032	1	753,411	0,000
Residuals	4117461	2243	-	-

Foram encontrados 433 pontos influentes mas não se encontrou justificativa plausível para exclusão dos mesmos. Por isso decidiu-se por manter essas observações já que seriam importantes na explicação da renda.

O fator de inflação de variância (Tabela 4) mostra que não existe multicolinearidade ( $VIF < 10$ ) sendo assim os coeficientes estão sendo estimados corretamente.

Tabela 4: Fator de inflação da variância do modelo ajustado

	GVIF	Df	$GVIF^{(1/(2*Df))}$
budget2	2,07	1	1,44
num_voted_users2	4,23	1	2,06
director_name2	3,83	34	1,02
genres2	4,14	30	1,02
num_user_for_reviews	3,72	1	1,93
duration	1,65	1	1,28
title_year	1,74	1	1,32
num_critic_for_reviews	3,42	1	1,85

O modelo escolhido explica aproximadamente 67,31% da variação da variável estudada (*gross2*). Entretanto a análise de resíduos (Figura 4) confirma a desconfiança de não atendimento dos pressupostos necessários para uma análise mais criteriosa dos resultados.

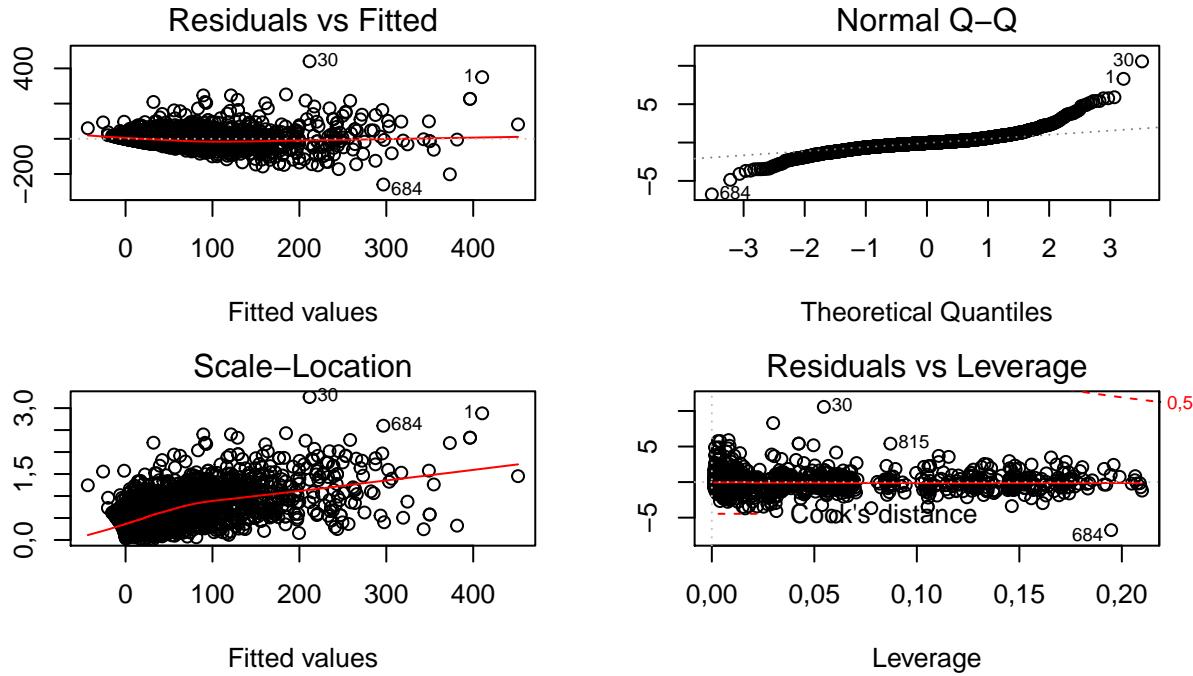


Figura 4: Análise gráfica do resíduo para modelo ajustado para com a variável resposta original

Utilizando o método analítico proposto por Box-Cox de transformação na variável resposta, encontra-se o parâmetro para realização da transformação potência. Entretanto, o intervalo de confiança contruído para o estimador de melhor expoente (Figura 5) não contem nenhum valor que permita transformações usuais, como log, raiz, inversa simples ou expoentes inteiros.

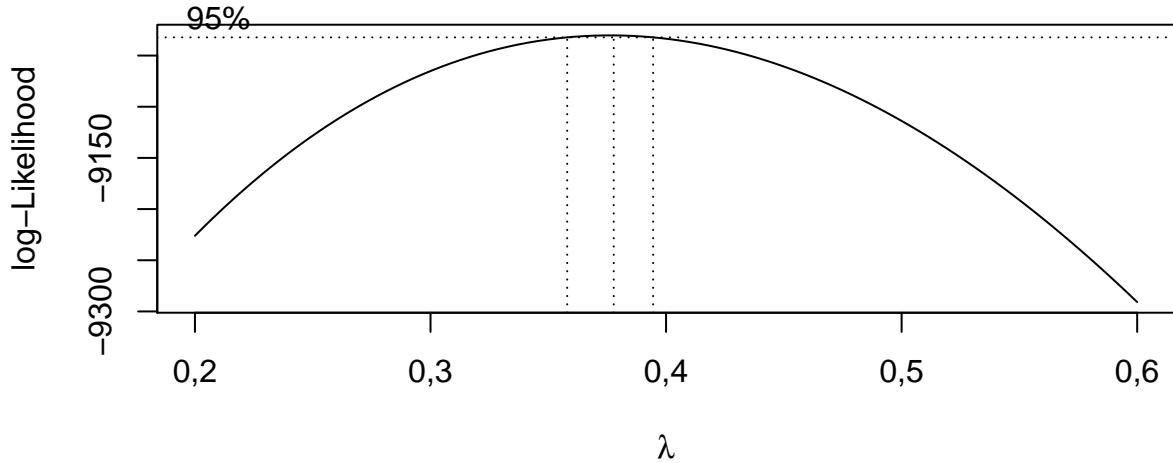


Figura 5: Intervalo de confiança de 95% para estimativa do expoente que maximiza o log da verossimilhança pelo método Box-Cox

Portanto, elevando a variável resposta (`gross`) a potência de 0,378 (ou  $\frac{189}{500}$ ) garante melhorias nos problemas de assimetria (falta de normalidade) e heterocedasticidade encontrado no valor da receita bruta original (Figura 6).

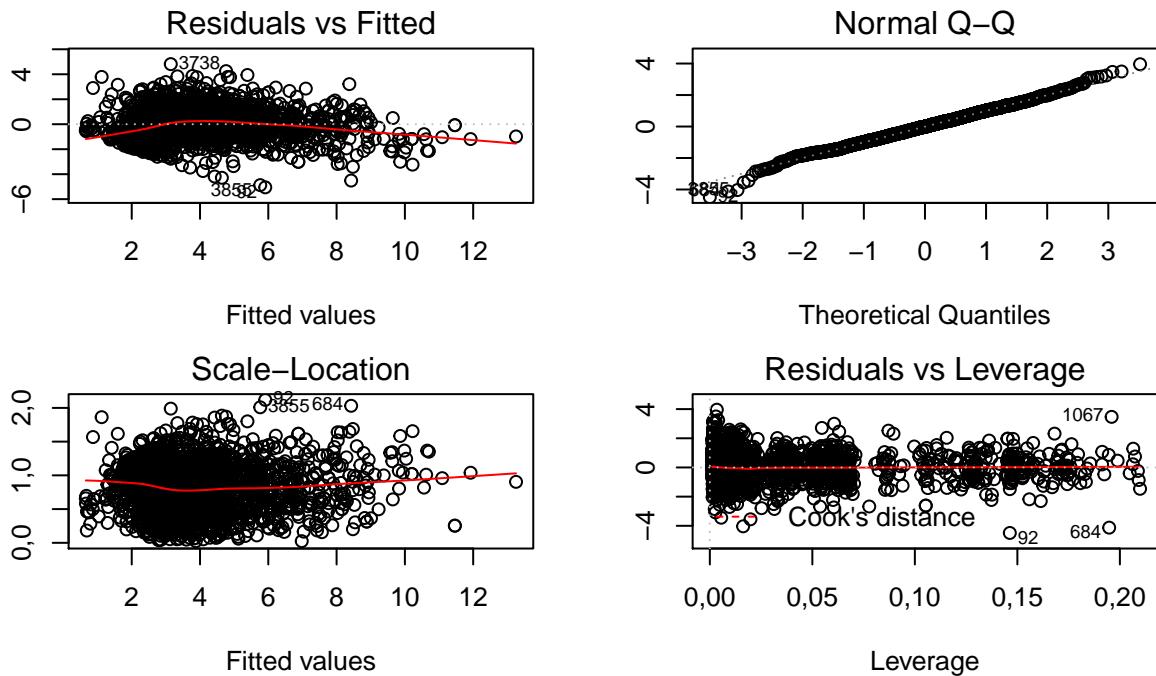


Figura 6: Análise gráfica do resíduo para modelo ajustado para com a variável resposta transformada

A variável transformada apresenta uma relação não linear com a original e portanto, a interpretação dos coeficientes também não será linear e como apresentado na Figura 7, quanto maior o valor da variável, maior será o efeito (seja positivo ou negativo) do regressor.

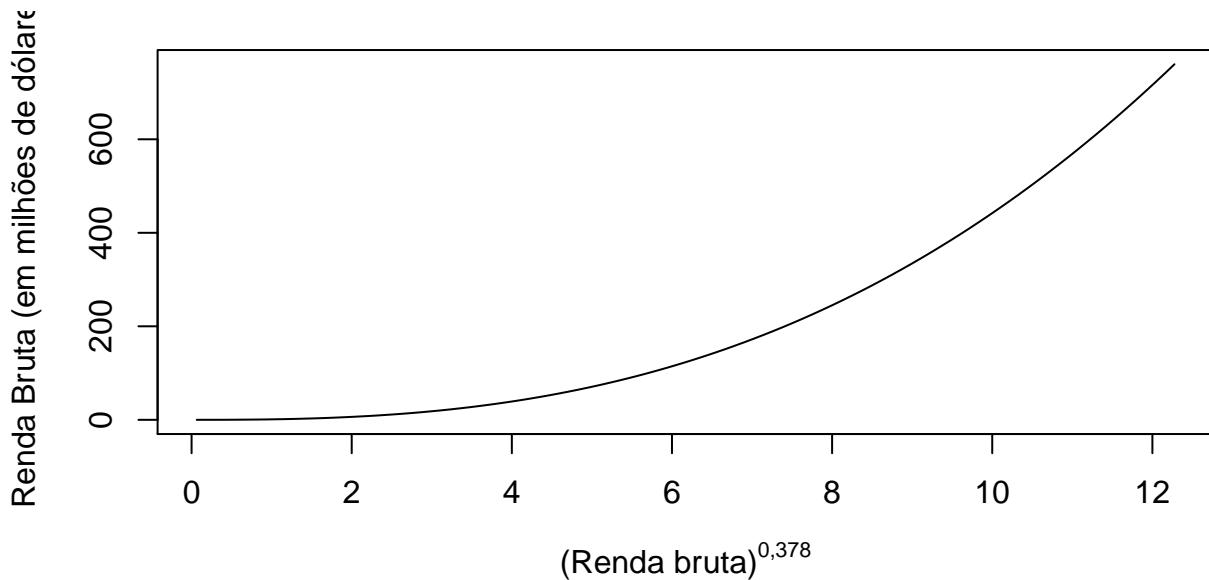


Figura 7: Correspondência da variável resposta renda bruta transformada com a escala original (em milhões de dólares)

## Conclusão

Utilizando os coeficientes expostos no Anexo e aplicando a inversa no resultado obtido pelo modelo é possível perceber um acréscimo de \$300 mil até aproximadamente \$3 milhões de dólares na média da renda bruta do filme para cada milhão de dólar no orçamento.

Acréscimos de aproximadamente \$100 mil a \$1 milhão; \$10 mil e de \$40 a \$80 mil na média da renda bruta são observados para cada acréscimo de unidade de votos dos usuários do IMDB, número de avaliações por usuários e número de avaliações por críticos profissionais, respectivamente.

Também pôde ser observado uma queda de aproximadamente \$330 mil na média da renda bruta por ano, no período avaliado e ainda uma queda de \$30 mil na média da renda para cada minuto a mais de filme.

Todos esses valores foram estimados mantendo as outras variáveis constantes.

Além disso, percebe-se diferenças na média da renda para cada diretor, destacando, de forma negativa, Christopher Nolan e Martin Scorsese e de forma positiva, Malcolm D. Lee e Tyler Perry.

Quanto às classes de gênero Comedy|Family se destacam positivamente enquanto Action|Adventure|Sci-Fi é a combinação de gêneros que rende menos para os filmes norte-americanos.

Futuros trabalhos considerando elenco de uma forma consistente e mesmo corrigindo as variáveis provenientes do Facebook podem vir a aprimorar a explicação de renda dos filmes norte-americanos nos últimos anos.

## Anexo

### Estimativas do modelo com variável resposta transformada

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	30,701	13,458	2,281	0,023
budget2	0,022	0,001	28,358	0,000
num_voted_users2	0,004	0,000	11,323	0,000
director_name2Antoine Fuqua	0,477	0,467	1,020	0,308
director_name2Bobby Farrelly	0,523	0,464	1,125	0,261
director_name2Brett Ratner	0,534	0,465	1,150	0,250
director_name2Catherine Hardwicke	0,582	0,502	1,158	0,247
director_name2Christopher Nolan	-3,496	0,476	-7,338	0,000
director_name2Clint Eastwood	0,444	0,359	1,238	0,216
director_name2David Ayer	-0,200	0,481	-0,416	0,677
director_name2David Fincher	-1,405	0,479	-2,930	0,003
director_name2David Gordon Green	-0,737	0,503	-1,466	0,143
director_name2Dennis Dugan	1,278	0,435	2,940	0,003
director_name2Garry Marshall	1,378	0,465	2,964	0,003
director_name2Gore Verbinski	-0,422	0,483	-0,875	0,382
director_name2James Mangold	0,228	0,466	0,490	0,624
director_name2James Wan	0,704	0,479	1,471	0,141
director_name2Kevin Smith	-0,409	0,465	-0,879	0,380
director_name2Malcolm D. Lee	1,515	0,549	2,760	0,006
director_name2Martin Scorsese	-1,808	0,476	-3,796	0,000
director_name2Michael Bay	0,292	0,399	0,731	0,465
director_name2M. Night Shyamalan	0,283	0,446	0,634	0,526
director_name2Renny Harlin	-0,105	0,466	-0,226	0,822
director_name2Ridley Scott	-1,136	0,394	-2,880	0,004
director_name2Rob Cohen	0,900	0,415	2,166	0,030
director_name2Robert Rodriguez	1,172	0,433	2,706	0,007
director_name2Robert Zemeckis	-0,697	0,465	-1,498	0,134
director_name2Ron Howard	0,054	0,413	0,131	0,896
director_name2Sam Raimi	-0,394	0,444	-0,887	0,375
director_name2Shawn Levy	1,059	0,355	2,985	0,003
director_name2Steven Soderbergh	0,322	0,345	0,934	0,350
director_name2Steven Spielberg	0,163	0,393	0,414	0,679
director_name2Tim Burton	0,161	0,393	0,410	0,682
director_name2Todd Phillips	1,146	0,472	2,427	0,015
director_name2Tyler Perry	2,426	0,506	4,798	0,000
director_name2Woody Allen	-1,014	0,437	-2,318	0,021
director_name2Zack Snyder	-0,828	0,442	-1,874	0,061
genres2Action Adventure Fantasy	-1,011	0,309	-3,274	0,001
genres2Action Adventure Sci-Fi	-1,239	0,217	-5,720	0,000
genres2Action Adventure Sci-Fi Thriller	-0,259	0,276	-0,940	0,347
genres2Action Comedy Crime	1,236	0,286	4,325	0,000
genres2Action Crime Drama Thriller	-0,278	0,235	-1,180	0,238

	Estimate	Std. Error	t value	Pr(> t )
genres2Action Crime Thriller	0,116	0,218	0,531	0,596
genres2Action Drama Thriller	-0,016	0,308	-0,052	0,959
genres2Adventure Animation Comedy Family	1,211	0,303	3,996	0,000
genres2Adventure Animation Comedy Family Fantasy	0,898	0,263	3,420	0,001
genres2Biography Drama	0,427	0,314	1,359	0,174
genres2Biography Drama Sport	0,516	0,302	1,707	0,088
genres2Comedy	0,269	0,125	2,142	0,032
genres2Comedy Crime	0,679	0,232	2,931	0,003
genres2Comedy Drama	-0,247	0,132	-1,869	0,062
genres2Comedy Drama Romance	0,020	0,130	0,151	0,880
genres2Comedy Family	1,655	0,319	5,188	0,000
genres2Comedy Romance	0,568	0,125	4,545	0,000
genres2Crime Drama Mystery Thriller	-0,038	0,275	-0,139	0,889
genres2Crime Drama Thriller	-0,228	0,175	-1,299	0,194
genres2Documentary	-1,191	0,312	-3,821	0,000
genres2Drama	-0,618	0,138	-4,466	0,000
genres2Drama Music	-0,234	0,299	-0,783	0,434
genres2Drama Mystery Thriller	0,151	0,321	0,470	0,638
genres2Drama Romance	-0,273	0,158	-1,728	0,084
genres2Drama Sport	-0,084	0,285	-0,296	0,767
genres2Drama Thriller	-0,192	0,245	-0,786	0,432
genres2Horror	0,437	0,228	1,919	0,055
genres2Horror Mystery	0,678	0,300	2,257	0,024
genres2Horror Mystery Thriller	0,567	0,250	2,273	0,023
genres2Horror Thriller	0,243	0,278	0,873	0,383
num_user_for_reviews	0,000	0,000	3,460	0,001
duration	-0,001	0,002	-0,747	0,455
title_year	-0,014	0,007	-2,129	0,033
num_critic_for_reviews	0,002	0,000	6,384	0,000