IMDB Data analysis

Donghao Xue

12/09/2020

1.Abstract

Every time before I watch a movie, I will check the rating score on IMBD which is a popular website providing source for movies. In this project, I choose a dataset that contains information of about five thousand movies from IMDB. I will mainly focus on finding the effect of different factors on the rating score of the movie. I plan to use two models to estimate the effect, one is linear regression model and the other one is multilevel model. At the end I found that those factors in the dataset have very weak relationship with the rating score. Therefore if we want to know the rating score will be effected by what kind of factors, we need to collect more information of movies.

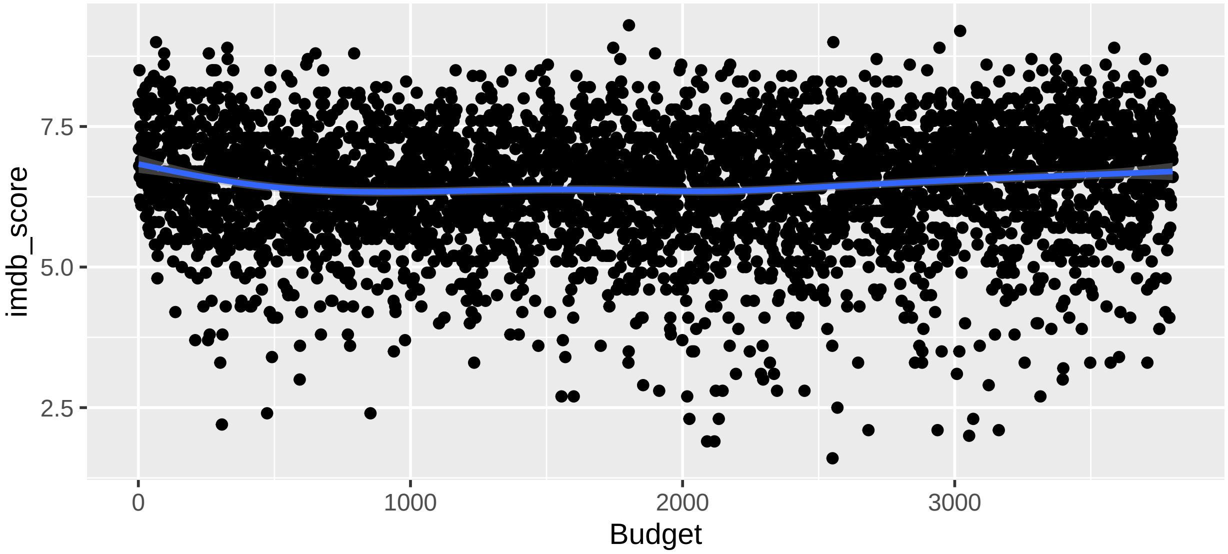
2.Introduction

This dataset is downloaded from the Kaggle website and it has 28 variables such as the country that movies are made, the rating score of movies(from 1 to 10), the number of critics for reviews, the duration of the movie and so on. I will consider the country, the duration of the movie, the number of critics for reviews, the budget as predictors, so that I can see how the rating score is effected by these predictors.

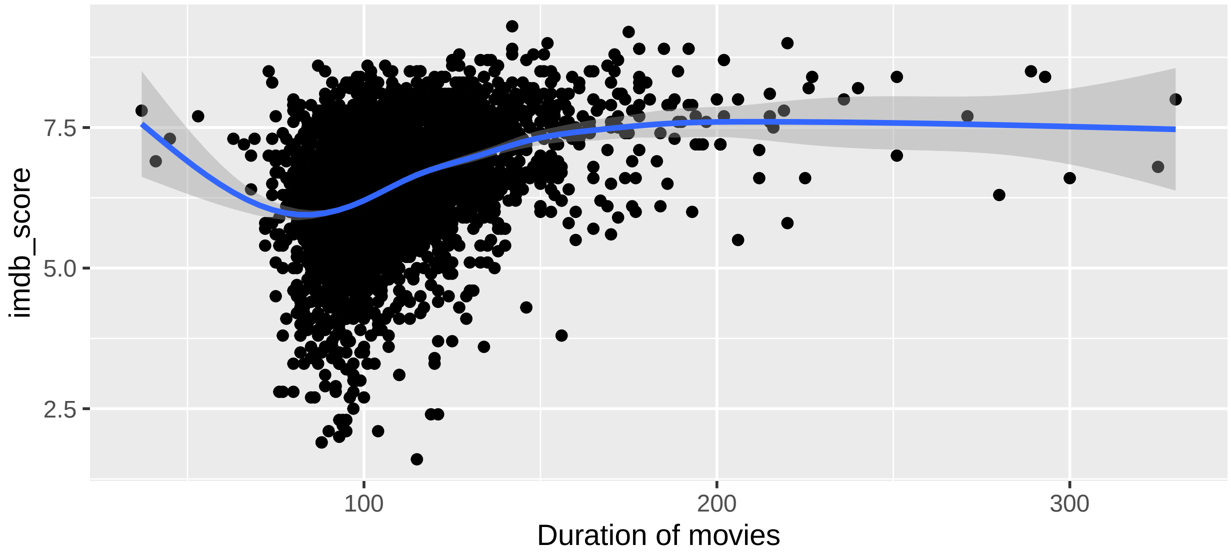
3.Method

3.1 EDA

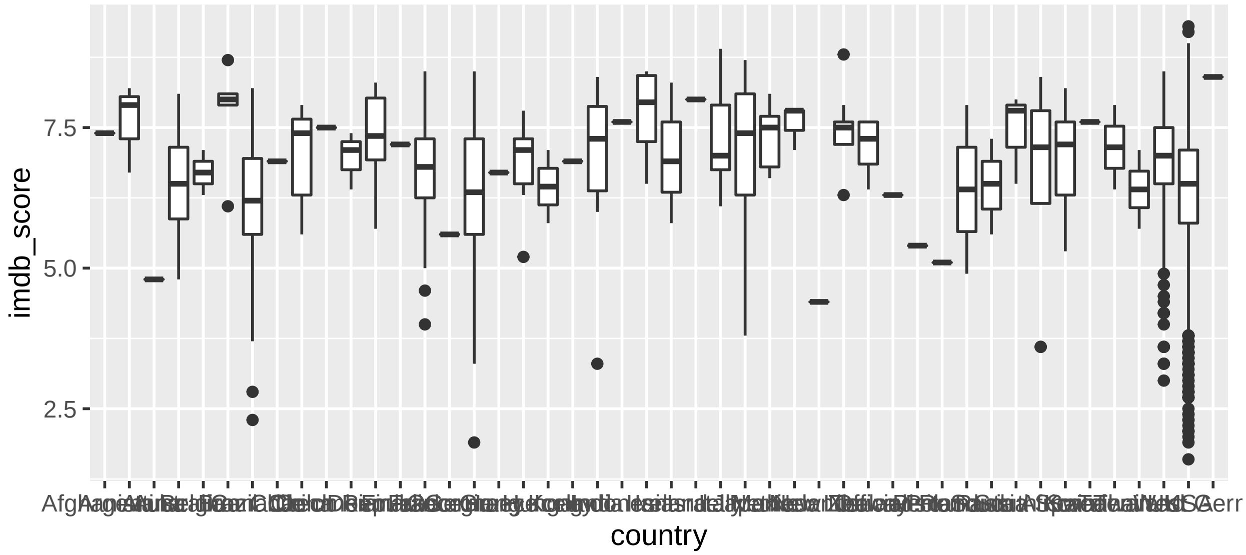
plot for rating score vs budget



plot for rating score vs duration of the movie



boxplot for rating score in each country



3.2 Model Fit

Model 1

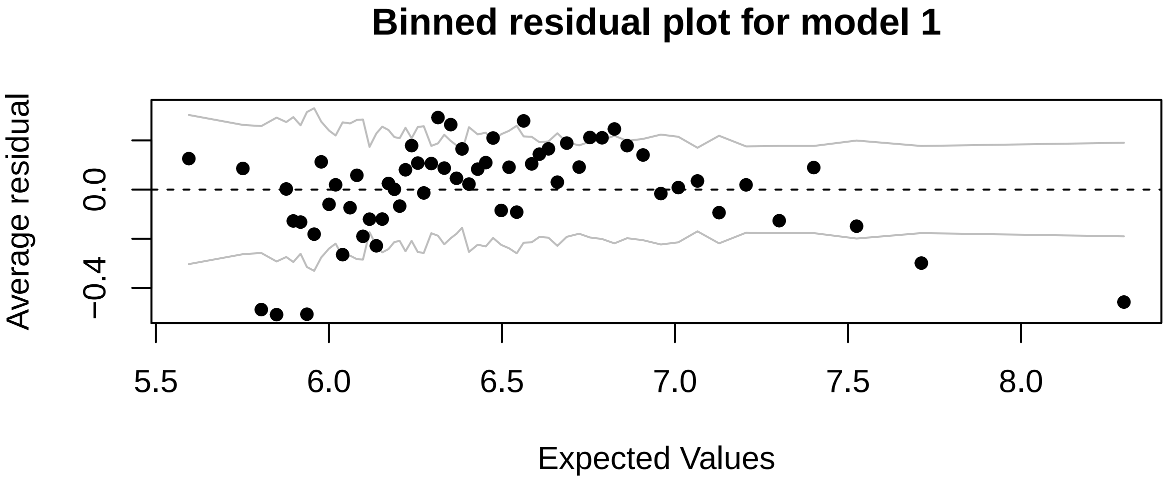
```
##
## Call:
## lm(formula = imdb_score ~ budget + duration + num_critic_for_reviews +
##     factor(country), data = imdb, refresh = 0)
##
## Coefficients:
##                    (Intercept)                                  budget
##                        6.00025                                -0.01069
##                       duration                  num_critic_for_reviews
##                       0.01389                       0.00235
##      factor(country)Argentina          factor(country)Aruba
##                      -0.35610                      -2.61811
##      factor(country)Australia        factor(country)Belgium
##                      -1.38051                      -0.91185
##         factor(country)Brazil         factor(country)Canada
##                      -0.13984                      -1.55063
##          factor(country)Chile          factor(country)China
##                      -1.14372                      -0.99756
##       factor(country)Colombia factor(country)Czech Republic
##                      -0.26722                      -0.96859
##        factor(country)Denmark        factor(country)Finland
##                      -0.76864                      -0.57348
##         factor(country)France        factor(country)Georgia
##                      -1.14710                      -2.14179
##        factor(country)Germany         factor(country)Greece
##                      -1.57001                      -1.47219
##      factor(country)Hong Kong        factor(country)Hungary
##                      -0.92754                      -1.29959
##        factor(country)Iceland          factor(country)India
##                      -0.49900                      -1.27816
##      factor(country)Indonesia           factor(country)Iran
##                      -0.94730                      -0.03851
##        factor(country)Ireland         factor(country)Israel
##                      -0.93049                       0.20686
##          factor(country)Italy          factor(country)Japan
##                      -0.81383                      -0.98546
##         factor(country)Mexico    factor(country)Netherlands
##                      -0.45765                      -0.45604
##       factor(country)New Line    factor(country)New Zealand
##                      -3.18108                      -1.07381
##         factor(country)Norway  factor(country)Official site
##                      -0.71436                      -1.64122
##           factor(country)Peru         factor(country)Poland
##                      -2.21285                      -2.36183
##        factor(country)Romania         factor(country)Russia
##                      -1.44334                      -1.60320
##   factor(country)South Africa    factor(country)South Korea
##                      -0.63200                      -1.35263
##          factor(country)Spain         factor(country)Sweden
##                      -0.92555                      -0.32511
##         factor(country)Taiwan       factor(country)Thailand
##                      -0.98951                      -1.90618
##             factor(country)UK            factor(country)USA
##                      -1.13126                      -1.52077
##   factor(country)West Germany
##                      -1.89469
```

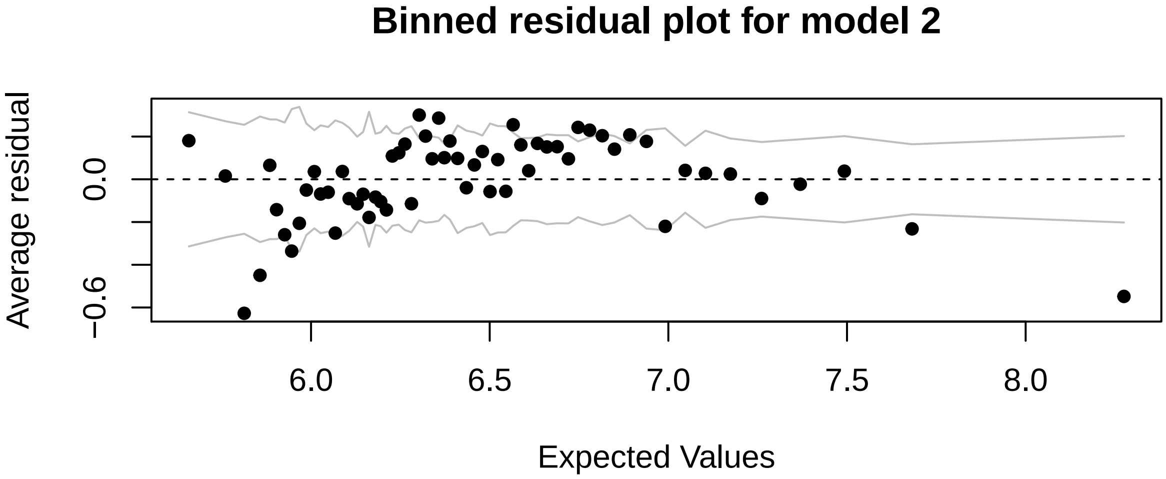
Model 2

```
## lmer(formula = imdb_score ~ budget + duration + num_critic_for_reviews + 
##     (1 | country), data = imdb)
##                         coef.est coef.se
## (Intercept)             4.8954   0.1047
## budget                  -0.0127   0.0071
## duration                 0.0138   0.0007
## num_critic_for_reviews  0.0024   0.0001
##
## Error terms:
##  Groups   Name        Std.Dev.
##  country  (Intercept) 0.2742
##  Residual             0.9283
## ---
## number of obs: 3801, groups: country, 46
## AIC = 10298.2, DIC = 10205.3
## deviance = 10245.8
```

4.Result

4.1 Model Check





4.2 Model Choose

Based on the binned residual plots for these two models, I think none of them is appropriate to fit the data since the average residuals does not have an regular pattern.

5.Discussion

5.1 Implication

From the coefficients of the model, we can see that the relationship between rating score and other predictors is not very strong. The predictors *duration* and *num_critic_for_reviews* have positive effects on the rating score and the predictor *budget* has an negative effect on the rating score. For the coefficients of different countries, the rating score of movies in Israel is higher than other countries relatively.

5.2 Limitation

It is hard to find information that has strong relationship to the rating score in the dataset.

5.3 Future Direction

I will try to collect more information about movie so that I can find out the relationship between rating score and other factors.

6.Reference

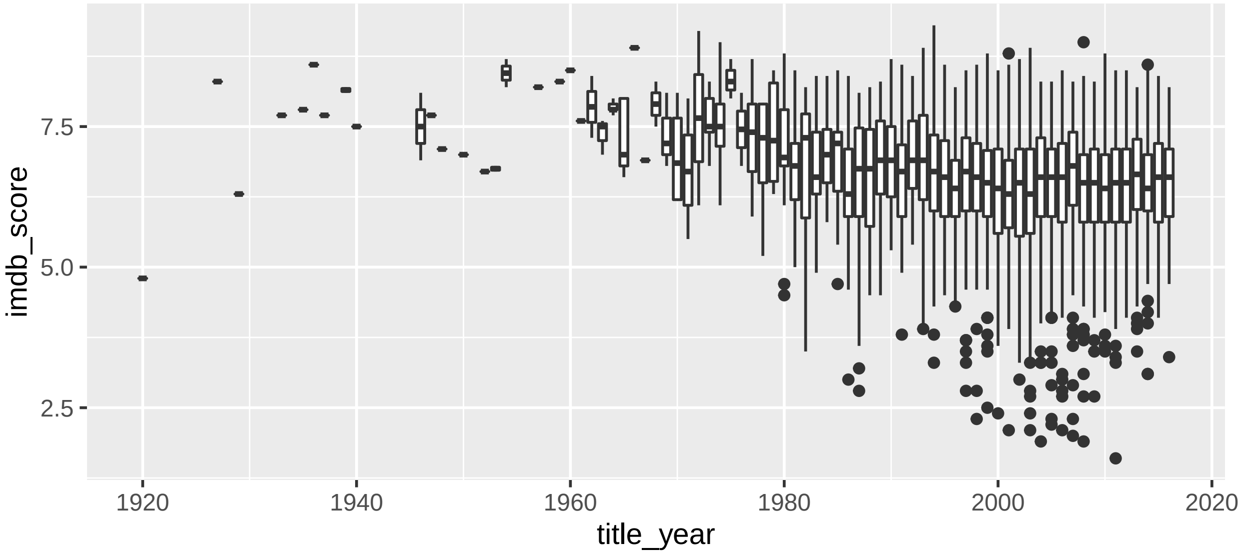
Goodrich, Ben; Gabry, Jonah; Ali, Iamd; Brilleman, Sam (2018). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.17.4, http://mc-stan.org/.

Hadley Wickham (2017). tidyverse: Easily Install and Load the tidyverse. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Data source: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

7.Appendix

boxplot for rating score in each year


imdb_score
title_year