Midterm Project Report

Runxin Yu

12/6/2019

Introduction

The dataset is called `Movie Industry: Three Decades of Movies` that contains 6820 movies in total(220 movies per year form 1986-2016) collected from 57 regions all over the world. The goal is to see how score of a movie is affected by different variables such as budget of the movie, revenue of the movie, duation of the movie, etc.

Columns of the Dataset

Each movie has the following attributes: * `budget`: the budget of a movie. Some movies don't have this, so it appears as 0 * `company`: the production company * `country`: country of origin * `director`: the director * `genre`: main genre of the movie * `gross`: revenue of the movie * `name`: name of the movie * `rating`: rating of the movie (R, PG, etc.) * `released`: release date (YYYY-MM-DD) * `runtime`: duration of the movie * `score`: IMDb user rating * `votes`: number of user votes * `star`: main actor/actress * `writer`: writer of the movie * `year`: year of release

Exploratory Data Analysis(EDA)

Data Cleaning

Reset the cells with a budget of 0 into NA and remove those rows from the data frame. New data frame without missing values is created in order for further exporation. Notice that some of the movies are produced by the same film company but are named differently, rename those companies so that their names are consistent. Another thing observed form the data is that there are 2179 film production companies, which is too large to set to be a random effect. Therefore, 7 of the world most famous film production companies,`Twentieth Century Fox`,`Columbia Pictures`,`Universal Pictures`,`Warner Bros.`,`Paramount Pictures`,`Walt Disney`, and `Metro-Goldwyn-Mayer`, as well as companies with counts greater that 50 are kept their original names, otherwise `others` is being labeled for the `company` column. This process reduce that number of different type of companies into 16, which can be used as a random effect.

```
##      budget                                          company
##  Min.   :     6000   Universal Pictures                  : 265
##  1st Qu.: 10000000   Warner Bros.                        : 257
##  Median : 23000000   Paramount Pictures                  : 222
##  Mean   : 36145602   Twentieth Century Fox Film Corporation: 174
##  3rd Qu.: 46000000   New Line Cinema                     : 145
##  Max.   :300000000   Columbia Pictures Corporation       : 137
##                      (Other)                             :3438
##       country                  director          genre          gross
##  USA      :3726   Woody Allen      :  30   Comedy   :1310   Min.   :      309
##  UK       : 366   Clint Eastwood   :  24   Action   :1099   1st Qu.:  6290905
##  France   : 108   Steven Soderbergh:  21   Drama    : 793   Median : 23455506
##  Germany  :  93   Steven Spielberg :  21   Crime    : 356   Mean   : 46074694
##  Canada   :  79   Ron Howard       :  20   Adventure: 291   3rd Qu.: 57782434
##  Australia:  36   Ridley Scott     :  19   Biography: 239   Max.   :936662225
```

```
##  (Other)  : 230   (Other)          :4503   (Other)  : 550
##                  name            rating           released       runtime
##  Bulletproof       :   2   R        :2247   2007-10-19:   8   Min.   : 69.0
##  Cocktail          :   2   PG-13    :1561   2012-04-20:   8   1st Qu.: 96.0
##  Deadfall          :   2   PG       : 659   2000-10-20:   7   Median :104.0
##  Death at a Funeral:   2   G        : 100   2002-10-11:   7   Mean   :107.6
##  Fair Game         :   2   NOT RATED:  38   2003-11-14:   7   3rd Qu.:117.0
##  Fantastic Four    :   2   UNRATED  :  20   2008-12-25:   7   Max.   :280.0
##  (Other)           :4626   (Other)  :  13   (Other)   :4594
##      score                      star          votes                  writer
##  Min.   :1.500   Nicolas Cage     :  38   Min.   :    183   Woody Allen :  29
##  1st Qu.:5.800   Bruce Willis     :  33   1st Qu.:  16110   Stephen King:  20
##  Median :6.400   Robert De Niro   :  32   Median :  43940   John Hughes :  14
##  Mean   :6.356   Denzel Washington:  31   Mean   :  95702   Luc Besson  :  13
##  3rd Qu.:7.100   Tom Hanks        :  31   3rd Qu.: 109393   Wes Craven  :  12
##  Max.   :9.300   Johnny Depp      :  30   Max.   :1861666   Joel Coen   :  11
##                  (Other)          :4443                     (Other)     :4539
##       year
##  Min.   :1986
##  1st Qu.:1996
##  Median :2003
##  Mean   :2002
##  3rd Qu.:2010
##  Max.   :2016
##
```

Visually Explore the Data

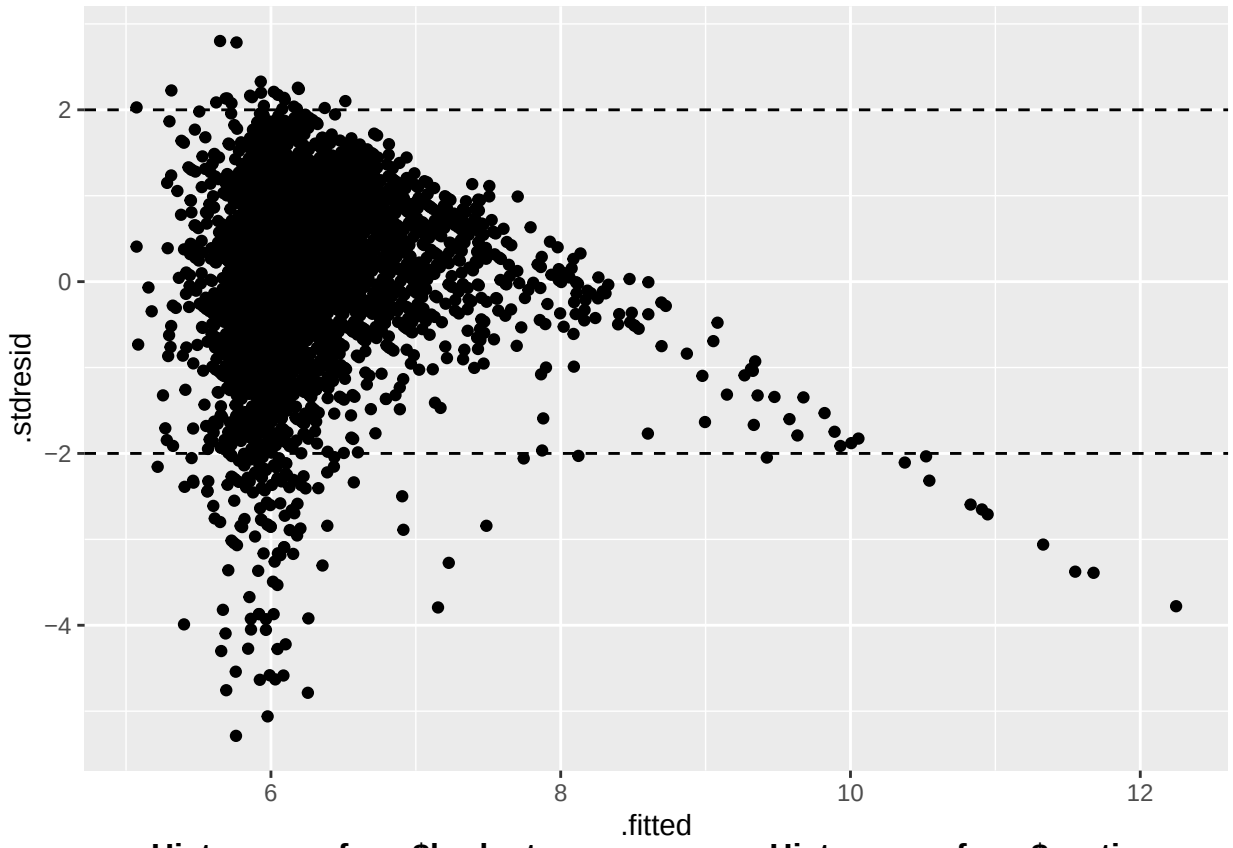
Though there are 57 regions in total in the data set, more than 80% of the data is collected from the United State. Therefore, the main focus is the film industries in the United State.

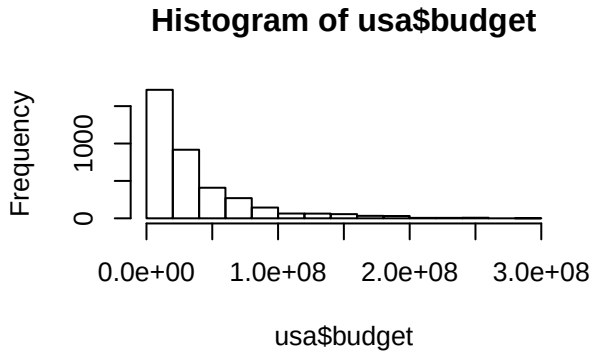
Linear Regression

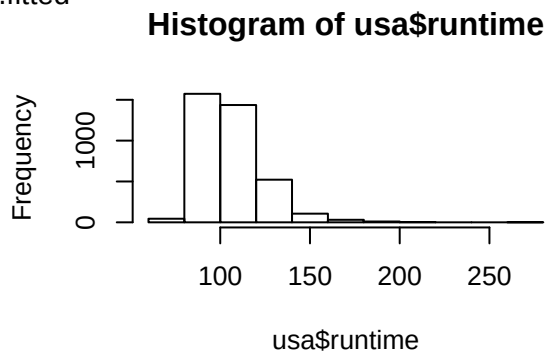
In order to see the relationships between ratings of movies and variables that may affect their scores furthermore, a simple linear regression is fitted to the original data.

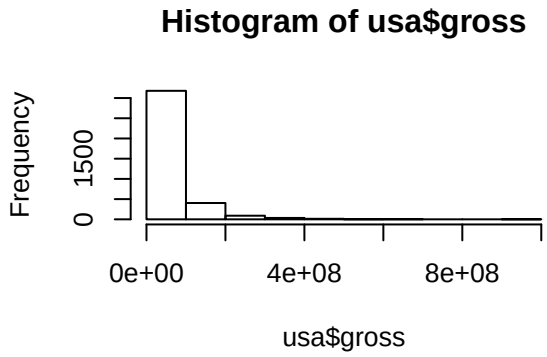
```
##
## Call:
## lm(formula = score ~ budget + runtime + gross + votes + factor(company),
##     data = usa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2589 -0.4012  0.0807  0.5327  2.2510
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           4.295e+00  1.036e-01  41.455  < 2e-16
## budget                               -6.143e-09  4.673e-10 -13.145  < 2e-16
## runtime                               1.746e-02  8.555e-04  20.405  < 2e-16
## gross                                 7.732e-10  3.088e-10   2.504 0.012326
## votes                                 2.986e-06  1.158e-07  25.789  < 2e-16
## factor(company)Dimension Films       -4.906e-01  1.297e-01  -3.782 0.000158
## factor(company)DreamWorks             1.493e-01  1.098e-01   1.360 0.173984
## factor(company)Fox 2000 Pictures      1.126e-02  1.261e-01   0.089 0.928813
```

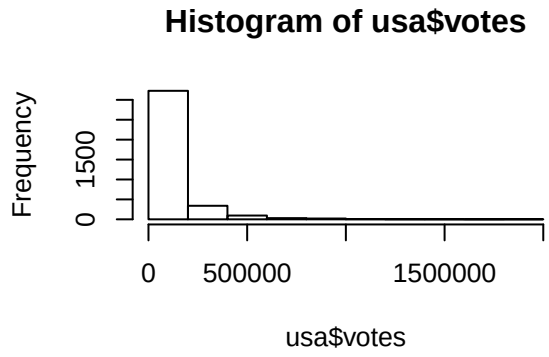
```
## factor(company)Lionsgate                 -2.433e-01  1.297e-01  -1.876 0.060738
## factor(company)Metro-Goldwyn-Mayer (MGM) -2.297e-02  1.105e-01  -0.208 0.835287
## factor(company)Miramax                    1.841e-01  1.277e-01   1.442 0.149443
## factor(company)New Line Cinema           -1.657e-01  9.002e-02  -1.840 0.065797
## factor(company)Others                     4.698e-02  5.755e-02   0.816 0.414394
## factor(company)Paramount Pictures         2.747e-02  7.824e-02   0.351 0.725489
## factor(company)Touchstone Pictures        1.493e-01  1.019e-01   1.465 0.142894
## factor(company)TriStar Pictures           1.835e-01  1.234e-01   1.486 0.137294
## factor(company)Twentieth Century Fox     -1.146e-01  8.237e-02  -1.391 0.164220
## factor(company)Universal Pictures         2.891e-02  7.656e-02   0.378 0.705715
## factor(company)Walt Disney                5.200e-01  9.778e-02   5.318 1.11e-07
## factor(company)Warner Bros.              -3.997e-02  7.546e-02  -0.530 0.596324
##
## (Intercept)                              ***
## budget                                   ***
## runtime                                  ***
## gross                                    *
## votes                                    ***
## factor(company)Dimension Films           ***
## factor(company)DreamWorks
## factor(company)Fox 2000 Pictures
## factor(company)Lionsgate                 .
## factor(company)Metro-Goldwyn-Mayer (MGM)
## factor(company)Miramax
## factor(company)New Line Cinema           .
## factor(company)Others
## factor(company)Paramount Pictures
## factor(company)Touchstone Pictures
## factor(company)TriStar Pictures
## factor(company)Twentieth Century Fox
## factor(company)Universal Pictures
## factor(company)Walt Disney               ***
## factor(company)Warner Bros.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.806 on 3706 degrees of freedom
## Multiple R-squared:  0.3527, Adjusted R-squared:  0.3494
## F-statistic: 106.3 on 19 and 3706 DF,  p-value: < 2.2e-16
```







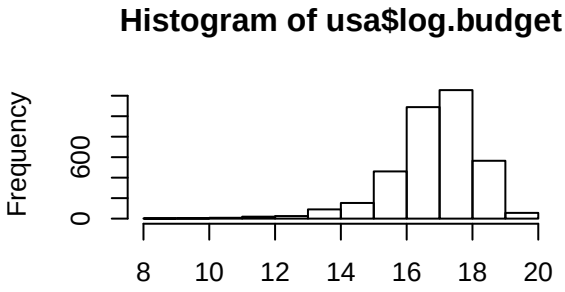


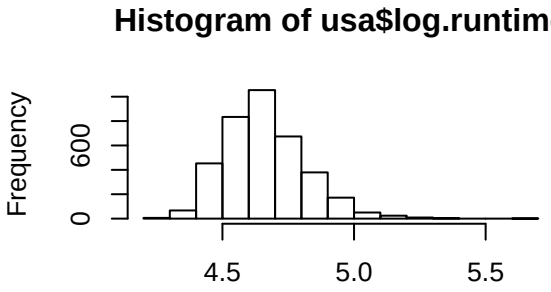


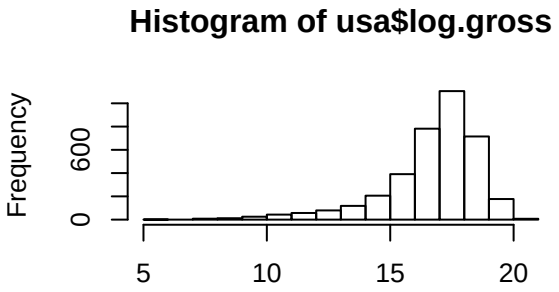
The residual plot generating form the simple linear regression shows both `Heteroscedasticity` and `Nonlinear` issue and the R-square of approximately 0.35 is not very promissing. This pattern indicates that transformation to some of the variables is needed. The histogram of variables with very large skewness also

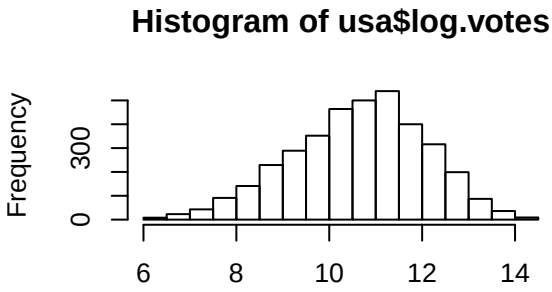
indicates that a log tansformation is needed.

Thus, log transformations are made to budget, runtime, gross and votes and fit the model again. The new model after log transformation is slightly better than the one one without log transformation



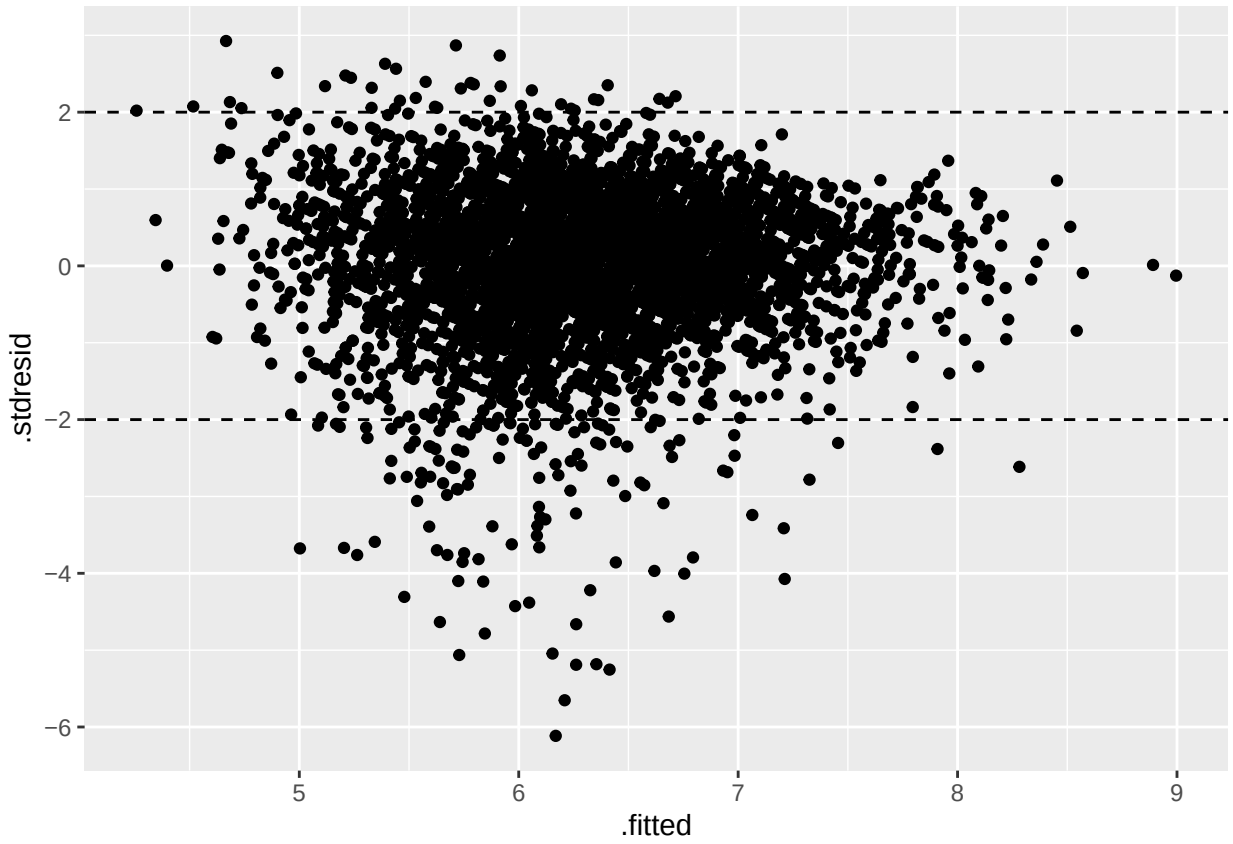







```
##
## Call:
## lm(formula = score ~ log.budget + log.runtime + log.gross + log.votes +
##     factor(company), data = usa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6686 -0.3902  0.0777  0.4949  2.2338
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -4.239175   0.404985 -10.467  < 2e-16
## log.budget                           -0.257291   0.012944 -19.877  < 2e-16
## log.runtime                           2.325874   0.092142  25.242  < 2e-16
## log.gross                            -0.015984   0.009302  -1.718 0.085808
## log.votes                             0.396036   0.011591  34.167  < 2e-16
## factor(company)Dimension Films       -0.471645   0.122996  -3.835 0.000128
## factor(company)DreamWorks             0.188629   0.104137   1.811 0.070166
## factor(company)Fox 2000 Pictures      0.114303   0.119533   0.956 0.339009
## factor(company)Lionsgate             -0.292396   0.122952  -2.378 0.017452
## factor(company)Metro-Goldwyn-Mayer (MGM)  0.155117   0.104926   1.478 0.139399
## factor(company)Miramax                0.284653   0.121070   2.351 0.018768
## factor(company)New Line Cinema       -0.127715   0.085264  -1.498 0.134250
## factor(company)Others                 0.105126   0.055286   1.901 0.057316
```

```
## factor(company)Paramount Pictures         0.114210   0.074196   1.539 0.123813
## factor(company)Touchstone Pictures        0.206587   0.096579   2.139 0.032498
## factor(company)TriStar Pictures           0.360482   0.117079   3.079 0.002092
## factor(company)Twentieth Century Fox     -0.046065   0.078109  -0.590 0.555389
## factor(company)Universal Pictures         0.055603   0.072595   0.766 0.443759
## factor(company)Walt Disney                0.528495   0.091842   5.754  9.4e-09
## factor(company)Warner Bros.               0.017742   0.071510   0.248 0.804069
##
## (Intercept)                              ***
## log.budget                               ***
## log.runtime                              ***
## log.gross                                .
## log.votes                                ***
## factor(company)Dimension Films           ***
## factor(company)DreamWorks                .
## factor(company)Fox 2000 Pictures
## factor(company)Lionsgate                 *
## factor(company)Metro-Goldwyn-Mayer (MGM)
## factor(company)Miramax                   *
## factor(company)New Line Cinema
## factor(company)Others                    .
## factor(company)Paramount Pictures
## factor(company)Touchstone Pictures       *
## factor(company)TriStar Pictures          **
## factor(company)Twentieth Century Fox
## factor(company)Universal Pictures
## factor(company)Walt Disney               ***
## factor(company)Warner Bros.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7646 on 3706 degrees of freedom
## Multiple R-squared:  0.4174, Adjusted R-squared:  0.4144
## F-statistic: 139.8 on 19 and 3706 DF,  p-value: < 2.2e-16
```


2
0
−2
−4
−6
.stdresid
5
6
7
8
9
.fitted