PSTAT 126 – Regression Analysis

Analysis of Movie Box-Office Grosses

Prof. Xiyue Liao

Arthur Li-Chuan Lee (Xu TR 5:00 PM) & Andy Tran (Xu TR 6:00 PM)

DUE September 14, 2018

Introduction

Our data set is the collection of movie releases in 2014 and various statistics collected via IMDb, YouTube, and Twitter. The goal of this report is to discover if gross revenue is affected by various predictors such as budget and IMDb ratings. We attempt to forecast the future gross profits of movies by using these predictors.

Questions of Interest

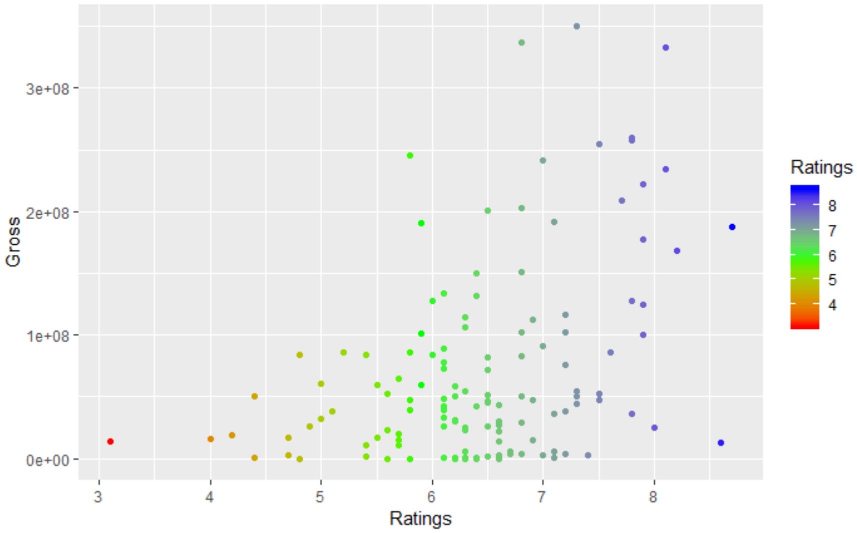
We consider the following research questions: Does a higher IMDb score contribute significantly to box office successfulness? Is budget a useful predictor of a movie's commercial success? Is a model containing at least one predictor from our data set useful in forecasting the gross revenue of a movie?

Regression Method

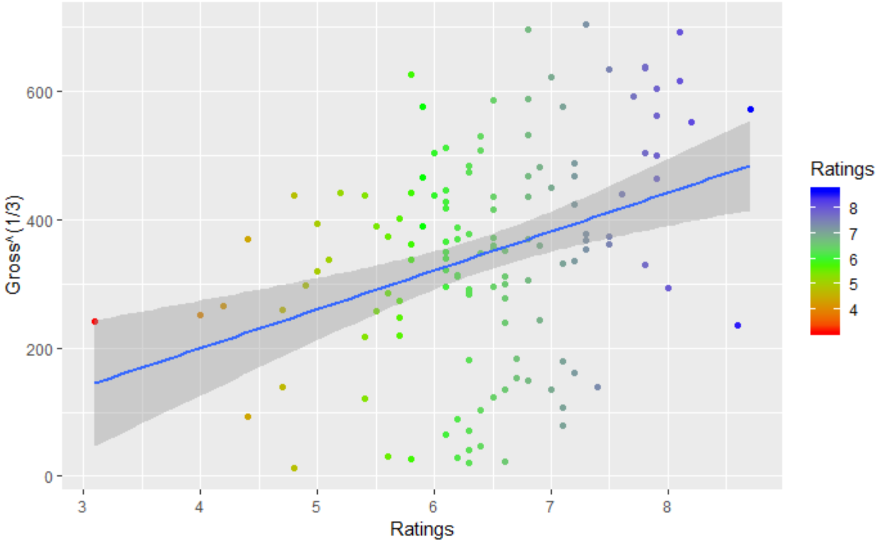
To answer our questions, we plot gross revenue of movies released in 2014 against their IMDb ratings and budgets, and use an information criterion to help us select the best model to use for predicting gross revenues of other movies released in other years.

Regression Analysis, Gross Revenue vs. IMDb Rating

An IMDb rating is a rating of a movie formed by the mean aggregate vote (from 1 to 10) of registered IMDb users. The scatterplot below shows gross revenue on IMDb ratings.



Regressing gross revenue on IMDb ratings results in a model that fails to meet the linearity, normality, and equal error variances criteria when observing the Residuals vs. Fitted and Normal Q-Q plots. To resolve these issues, we employ a Box-Cox power transformation and discover that the best regression is a regression of $Y^{1/3}$ on x.



From the plot above, we see that a linear relationship exists between transformed gross revenue and IMDb rating. The transformation significantly improves the Residual vs. Fitted and Normal Q-Q plots to better fulfill the LINE conditions. We use this transformed model to answer the following question.

Research Question: Are higher IMDb ratings an indicator of higher gross revenue among films?

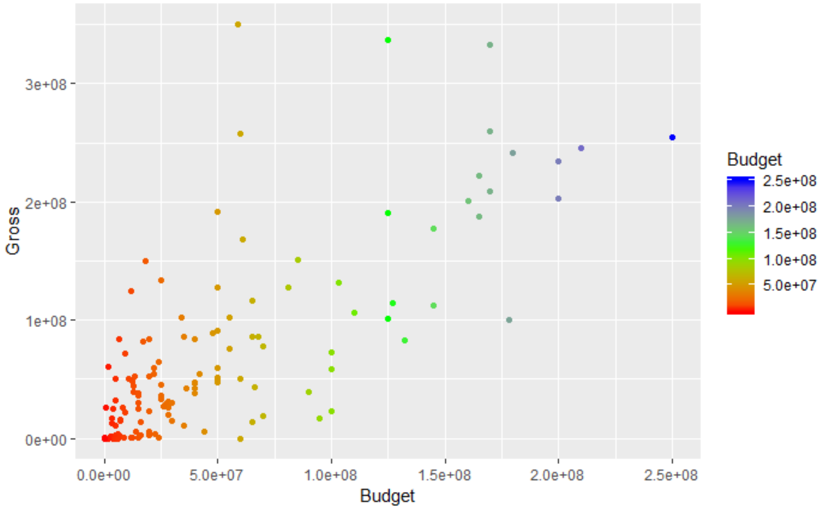
Yes. We conduct the following hypothesis test:

$$H_o: \beta_{Ratings} = 0 \ vs. \beta_{Ratings} \neq 0$$

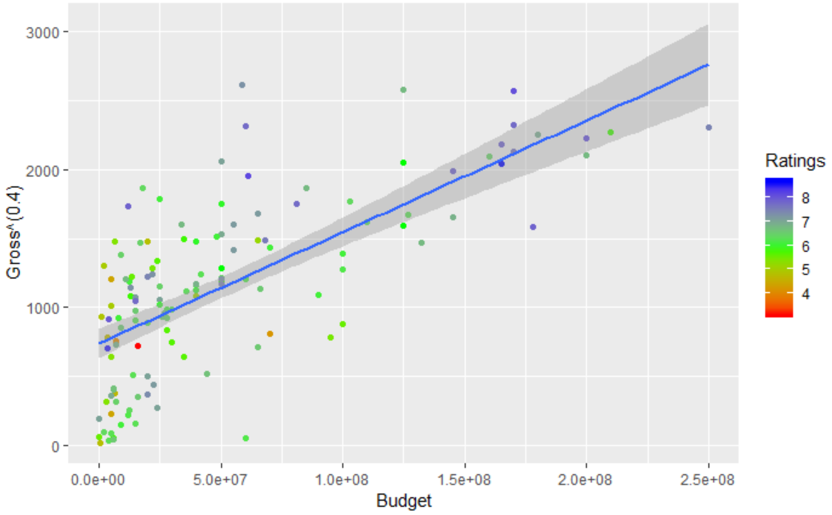
We execute a student's *t*-test and obtain a *t*-value of 4.225 corresponding to a p-value < 0.001. This enables us to reject the null hypothesis of our test at the $\alpha = 0.05$ significance level and conclude that IMDb ratings have a statistically significant effect on gross revenue.

Regression Analysis, Gross Revenue vs. Budget

Below is a scatterplot of gross revenue on movie budget. The plot of the data is fanning outwards which indicates nonconstant error variance. However, there are films with medial or low budgets that have relatively high gross revenues.



The Residual vs. Fitted and Normal Q-Q plots possess nonconstant error variance and non-normality issues. This is consistent with the fanning behavior we noted earlier. The issues are solved by using Box-Cox power transformation similar to the one used previously. This time, the best power to use is 0.4, so the ideal regression to use is a regression of $Y^{0.4}$ on x.



Observing the Residual vs. Fitted plot and Normal Q-Q plot for the transformed regression reveals that the transformation now fulfills the LINE criteria since the nonconstant error variance and non-normality problems have been resolved. We see from the graph above there is a linear relationship between gross revenue to the power of 0.4 and movie budget. In this model we see that 48.8% of the variation in gross revenue is explained in the predictor budget. With this new regression we can now answer the following research question.

Research Question: Is movie budget significantly related to box office success of films?

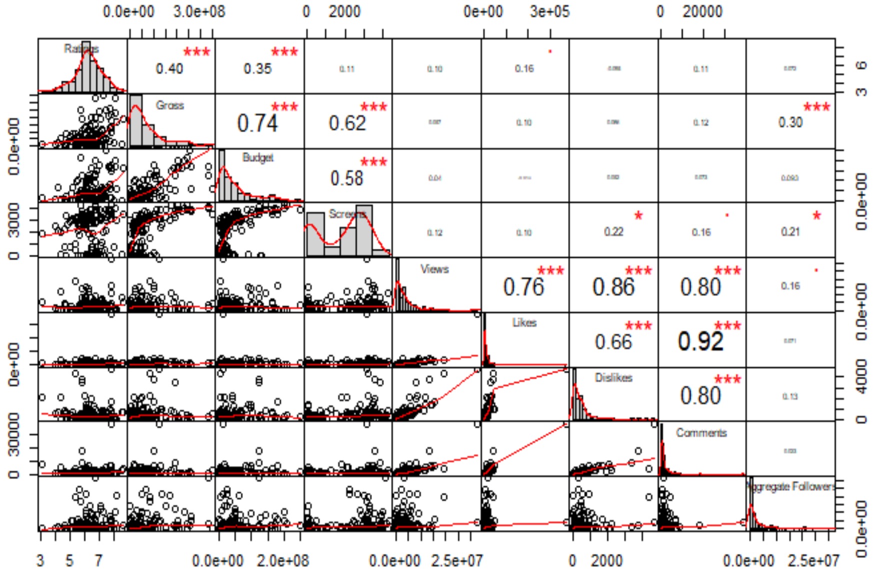
Yes. We conduct the following hypothesis test:

$$H_o: \beta_{Budget} = 0 \; vs. H_a: \beta_{Budget} \neq 0$$

We obtain a *t*-statistic of 11.21, which has a p-value < 0.001, from a *t*-test. From our *t*-test we obtained a *t*-statistic of 12.551 with a p-value < 0.001. Thus we reject the null hypothesis at the α = 0.05 significance level and conclude that budget has a statistically significant effect on box office success.

Regression Analysis, Building a Model for Gross Revenue

In scatterplot matrix below, we attempt to search for any correlation between gross revenue and other predictors within the dataset such as YouTube likes, dislikes, and views, aggregate followers on social media, and number of movie screenings.



Observations from the Scatterplot Matrix:

- Gross revenue and budget are highly correlated, which is consistent with the result of our *t*-test earlier in the report.
- There is a slight correlation between gross revenue and social media following of 0.30.
- The values of gross revenue on ratings are heavily skewed towards the middle.
- Gross revenue and amount of movie screenings are very correlated (0.58).
- Budget, screenings, aggregate followers, and ratings are statistically significant***.

Building a Model to Predict Gross Revenue:

We use Akaike's Information Criterion to select predictors for the final model. The criterion factors a model based on movie budget, amount of movie screenings, followers on social media, and movie ratings on IMDb. All insignificant factors with a p-value less than α = 0.05 were removed from the final model.

Initially, our model fails to meet the normality and equal error variances criteria. To solve these problems, we once again employ the Box-Cox power transformation to find the ideal regression to be a regression of $Y^{0.4}$ on x. The R^2 value of this fit is 0.685 which is significantly better than the R^2 value of the regression of log(Y) on log(x's) fit, 0.585, which we also attempted. When we check our new Residuals vs. Fitted and Normal Q-Q plots, we can see that the Residuals vs. Fitted plot now possesses a well-behaved pattern and that there is marked improvement in the normality of the data. Our transformed data now meets the LINE criteria and allows us to answer the following question.

Research Question: Is a model containing at least one predictor from our data set useful in forecasting the gross revenue of a movie?

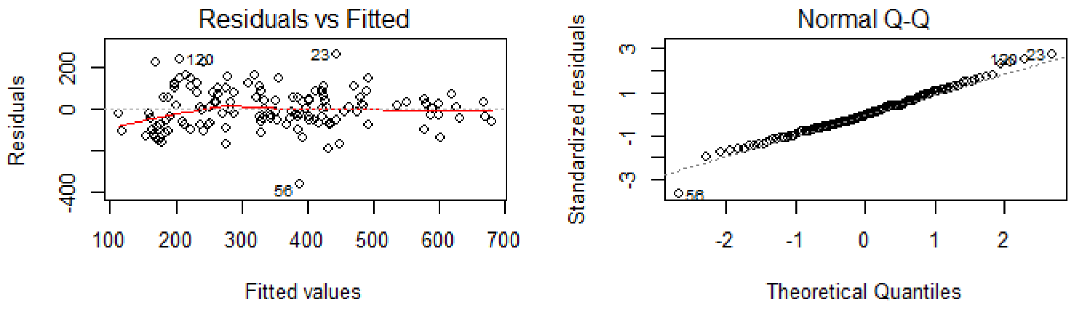
Yes. We conduct the following hypothesis test to answer our question of interest:

$$\beta_{Ratings} = \beta_{Budget} = \beta_{Screens} = \beta_{Followers} = 0 \; vs. At\ least\ one\ \beta_k \neq 0$$

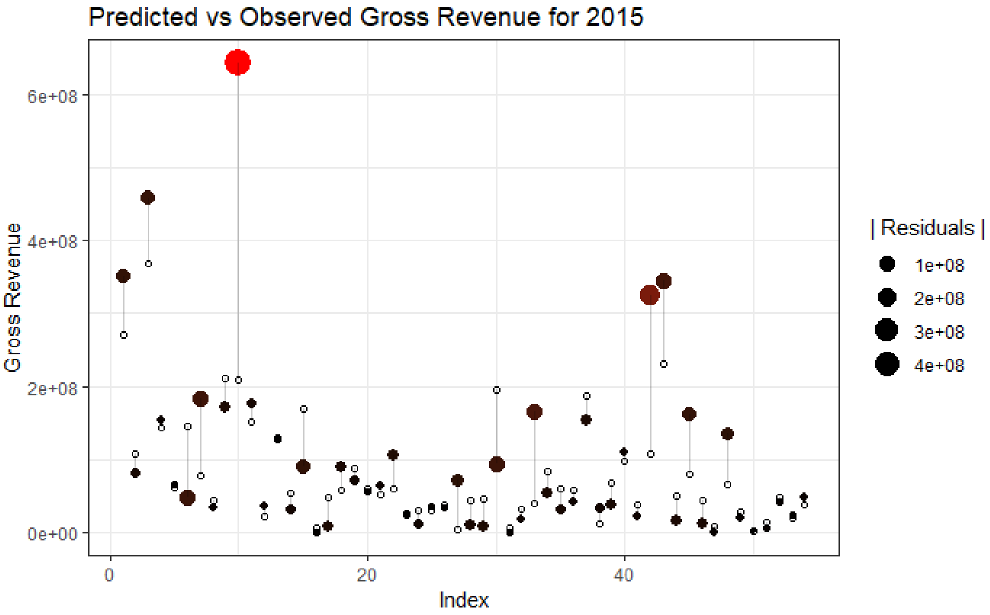
We conduct the test with a general linear F-test, the full model being $Y_i = \beta_o + \beta_{Ratings}\, x_{i,Ratings} + \beta_{Budget}\, x_{i,Budget} + \beta_{Screens}\, x_{i,Screens} + \beta_{Followers}\, x_{i,Followers} + \varepsilon_i$ and the reduced model being $Y_i = \beta_o + \varepsilon_i$. The F-statistic for the test is 12.39 and its p-value < 0.001. Thus, we reject the null hypothesis and conclude that at least one predictor must be useful in extrapolating the gross revenue of a movie.

Residual Analysis of the Final Model:

Our final model does not meet the LINE criteria. We use a Box-Cox transformation again to find a better regression model to fix the issues. Residual analysis of the regression model of $Y^{0.4}$ on the x's shows us that the normality and variability issues are fixed by this transformed model.



We now use the refined final model to predict revenue for movies released in 2015. Below is the plot of residuals – the actual gross revenues of movies released in 2015 in comparison to the corresponding predictions.



There is one glaring outlier in our new plot – Jurassic World. Our model projected that Jurassic World would make $209,656,741 in gross revenue. Instead, the movie was a Hollywood blockbuster and collected in excess of $600M in revenue. However, this isn't a particularly surprising result when you account for the fact Jurassic World is the long awaited reboot of the classic Jurassic Park series of which it had been 14 years since a movie from the franchise had been released, something our model simply could not represent. When the mixed critical reception it received is also factored in, our model was further fooled into severely underestimating Jurassic World's gross revenue by its heavy reliance on the 'Ratings' predictor ($\beta_{Ratings}$ = 115.3). Aside from this particular situation, our model was relatively accurate in projecting gross revenue accounting for 67.55% variation for all predictors.

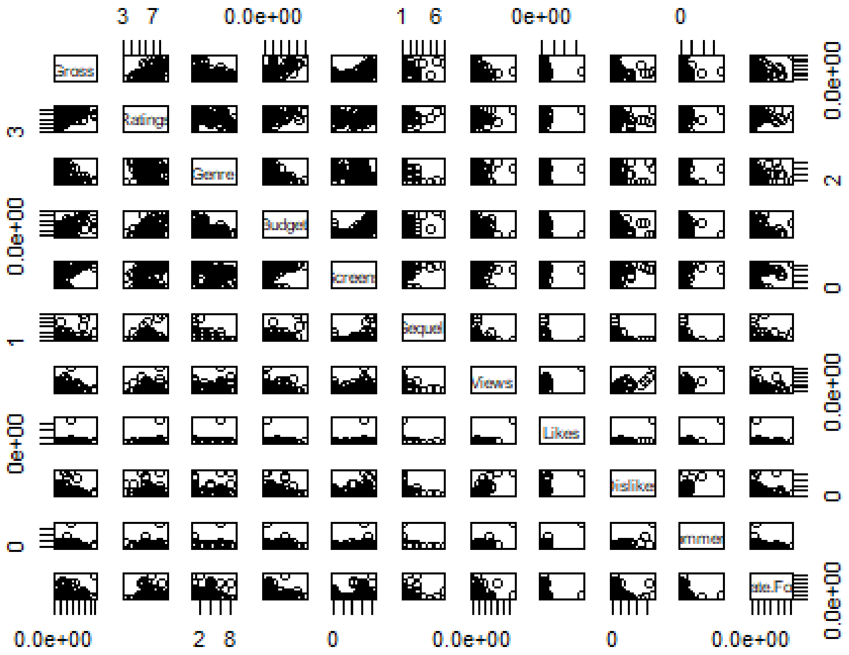
Conclusion

In summary, we found several predictors that were influential to a movie's success in the box office. IMDb rating was shown to have a positive correlation on the effect of movie revenue. We saw that larger movie budgets corresponded to better box office revenue. The number of screenings was also a factor in box office success. However, social media statistics such as YouTube views, likes, and dislikes were not strong indicators of box office success or failure, and movies with large social media followings had moderate correlation with box office success.

A larger sample size of movies can be added to the dataset for improvement of the final model. Other predictors like what kind of actors are in the film, which company produced the film, and amount spent on advertisements could also improve the model. Overall our results from our final model are very rigid, as seen from our predicted values for movies in 2015.

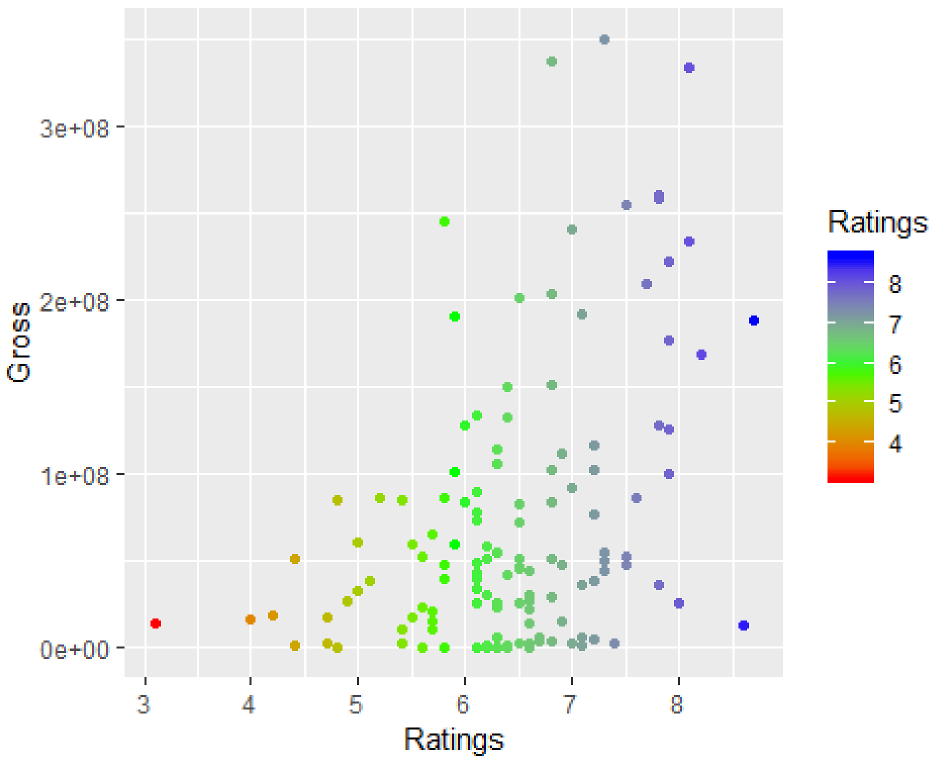
Appendix

```r
# Initialization of Variables for 2014 Movies, Omitting Data with N/A Values
CSM2014naomit = na.omit(CSM2014)
Ratings = CSM2014naomit$Ratings
Genre = factor(CSM2014naomit$Genre)
Budget = CSM2014naomit$Budget
Screens = CSM2014naomit$Screens
Sequel = CSM2014naomit$Sequel
Sentiment = CSM2014naomit$Sentiment
Views = CSM2014naomit$Views
Likes = CSM2014naomit$Likes
Dislikes = CSM2014naomit$Dislikes
Comments = CSM2014naomit$Comments
Aggregate.Followers = CSM2014naomit$Aggregate.Followers
Gross = CSM2014naomit$Gross
# Scatterplot Matrix of Gross vs. All Predictors, A First Peek at the Data
pairs(Gross~Ratings+Genre+Budget+Screens+Sequel+Views+Likes+Dislikes+Comments
+Aggregate.Followers)
```

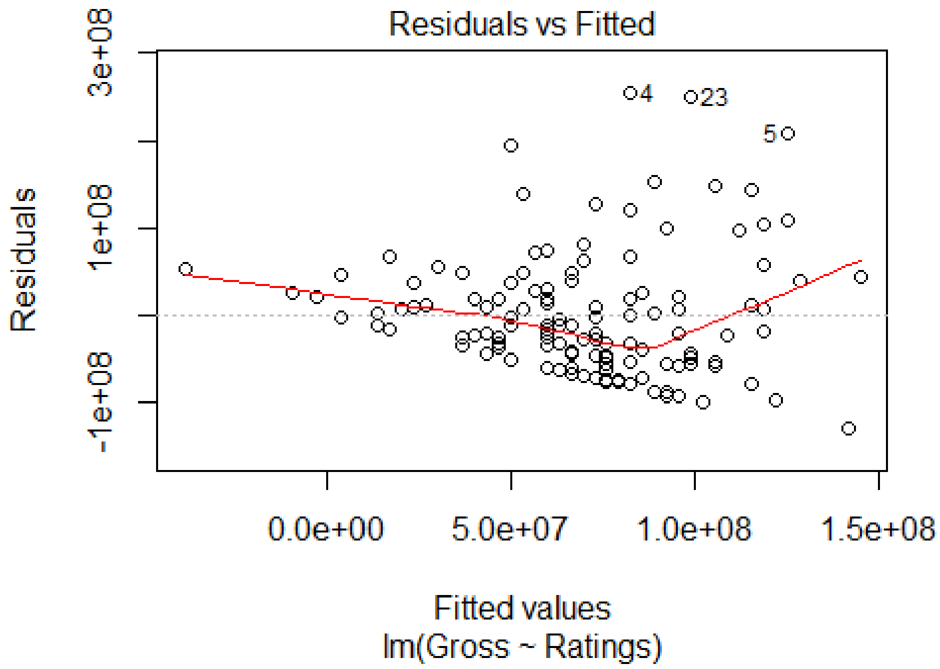


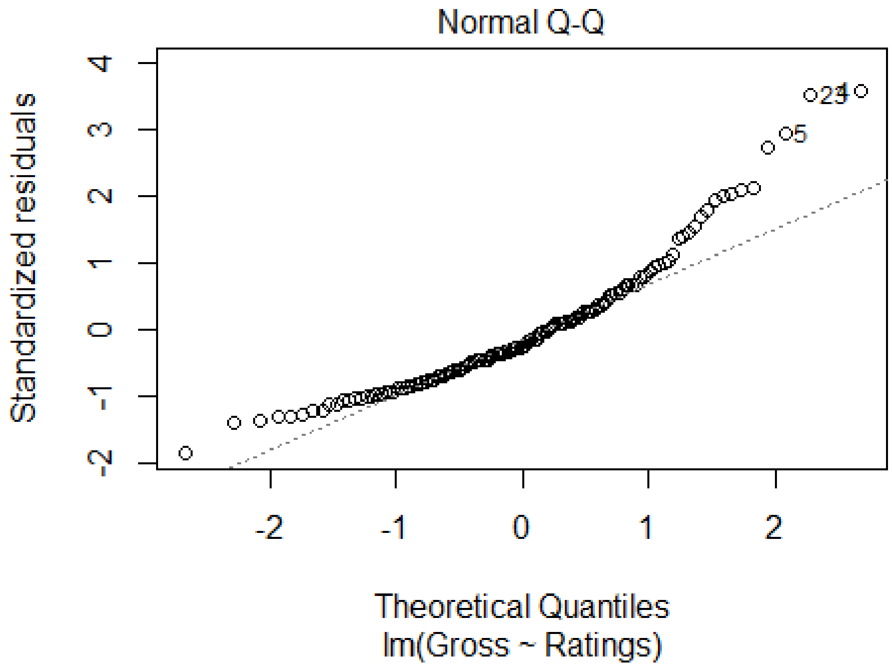
```r
# Regression of Gross on All Predictors
CSM2014LoBF = lm(Gross~Ratings+Genre+Budget+Screens+Sequel+Views+Likes+Dislik
es+Comments+Aggregate.Followers)
summary(CSM2014LoBF)
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Genre + Budget + Screens + Sequel +
##     Views + Likes + Dislikes + Comments + Aggregate.Followers)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -80774242 -25740509  -6138646  19153438 246972854
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.125e+08  3.625e+07  -3.104  0.00240 **
## Ratings              2.012e+07  5.368e+06   3.749  0.00028 ***
## GenreAdventure       1.080e+07  1.727e+07   0.625  0.53308
## GenreAnimation      -5.031e+06  1.734e+07  -0.290  0.77220
## GenreComedy         -1.152e+07  1.372e+07  -0.840  0.40273
## GenreCrime          -2.966e+07  1.885e+07  -1.574  0.11829
## GenreDrama          -1.823e+07  1.381e+07  -1.320  0.18943
## GenreHistory        -1.867e+07  1.833e+07  -1.019  0.31042
## GenreHorror          2.157e+07  2.577e+07   0.837  0.40436
## GenreOther          -1.191e+07  3.409e+07  -0.350  0.72733
## Budget               6.620e-01  1.220e-01   5.428 3.22e-07 ***
## Screens              1.179e+04  3.562e+03   3.311  0.00124 **
## Sequel               1.034e+06  5.082e+06   0.203  0.83916
## Views               -6.243e+00  2.217e+00  -2.816  0.00572 **
## Likes                7.915e+02  3.876e+02   2.042  0.04342 *
## Dislikes             2.108e+04  1.428e+04   1.476  0.14268
## Comments            -3.591e+03  3.939e+03  -0.912  0.36380
## Aggregate.Followers  2.515e+00  7.911e-01   3.179  0.00190 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45140000 on 115 degrees of freedom
## Multiple R-squared:  0.709,  Adjusted R-squared:  0.666
## F-statistic: 16.48 on 17 and 115 DF,  p-value: < 2.2e-16

# Scatterplot and Regression of Gross Revenue vs. Ratings
ggplot(CSM2014naomit, aes(x=Ratings, y=Gross, color=Ratings))+geom_point()+sc
ale_color_gradientn(colours=rainbow(3))
```

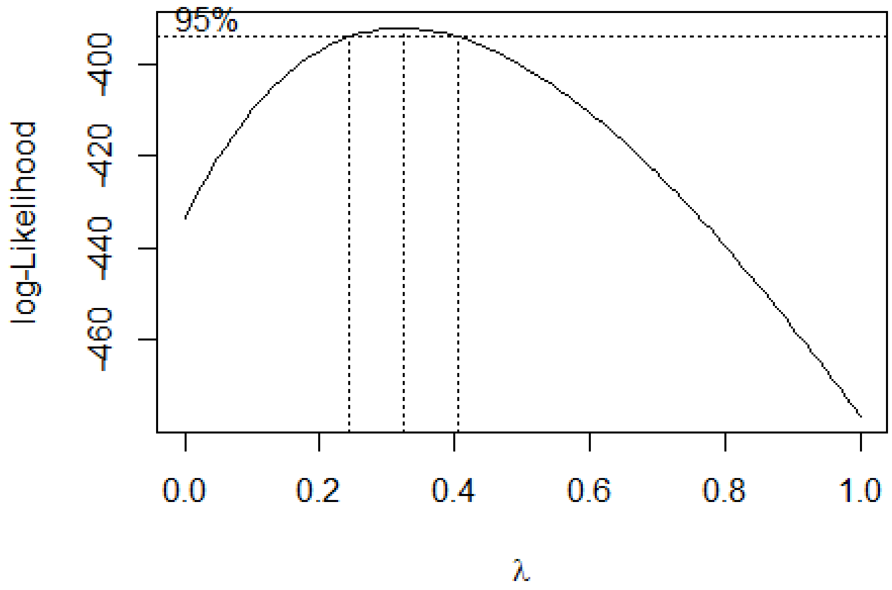


```
CSM2014LoBF1 = lm(Gross~Ratings)
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014LoBF1, which=c(1,2))
```

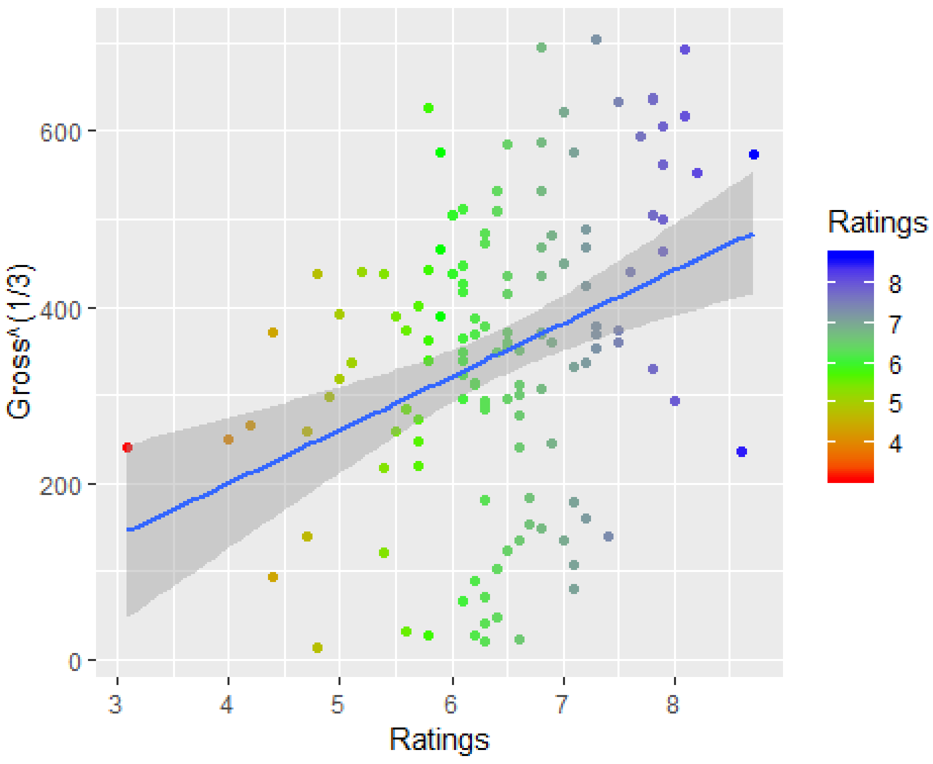




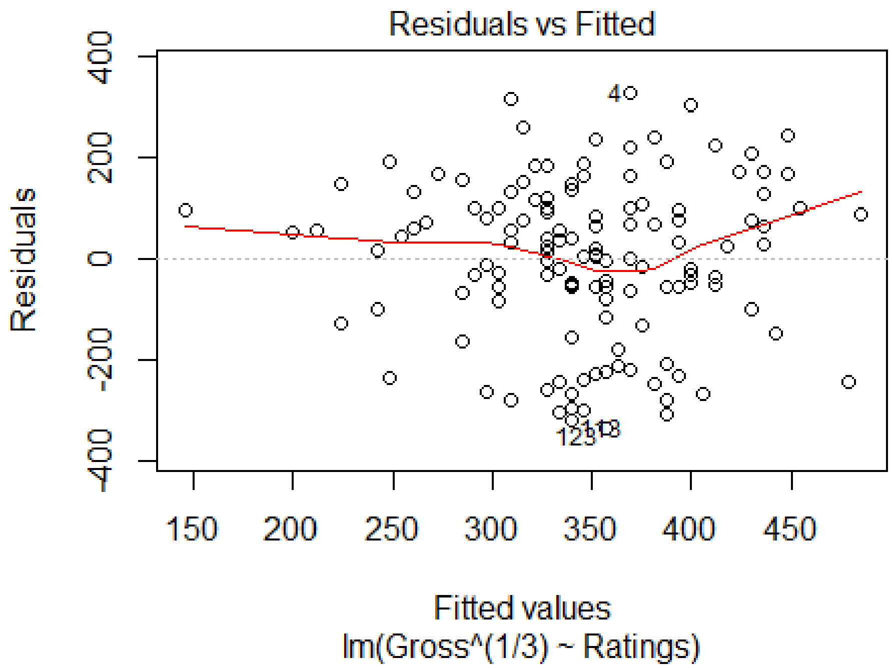
```
# Box-Cox Power Transformation to Find the Optimal Transformed Response Power
boxcox(Gross~Ratings, lambda=seq(0,1,0.1))
```

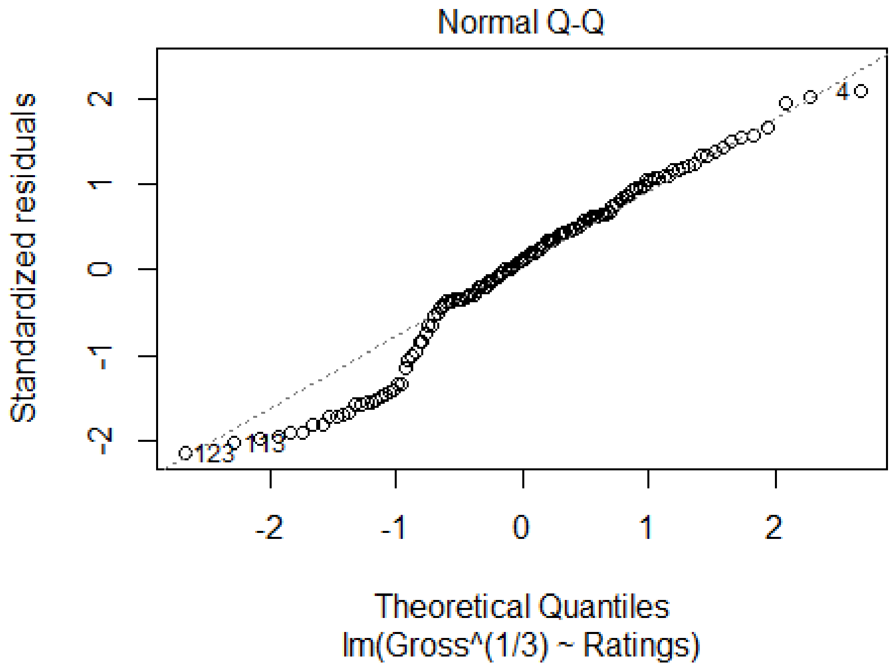


```
# Regression with Cube Root Response
CSM2014boxcox1 = lm(Gross^(1/3)~Ratings)
summary(CSM2014boxcox1)

##
## Call:
## lm(formula = Gross^(1/3) ~ Ratings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -334.69  -78.96   19.93  100.09  326.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -42.26      93.27  -0.453    0.651
## Ratings        60.56      14.34   4.225 4.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 157.8 on 131 degrees of freedom
## Multiple R-squared:  0.1199, Adjusted R-squared:  0.1132
## F-statistic: 17.85 on 1 and 131 DF,  p-value: 4.452e-05

# Scatterplot and Regression of Cube Root of Gross Revenue vs. Ratings
ggplot(CSM2014naomit, aes(x=Ratings, y=Gross^(1/3), color=Ratings))+geom_poin
t()+scale_color_gradientn(colours=rainbow(3))+geom_smooth(method=lm)
```

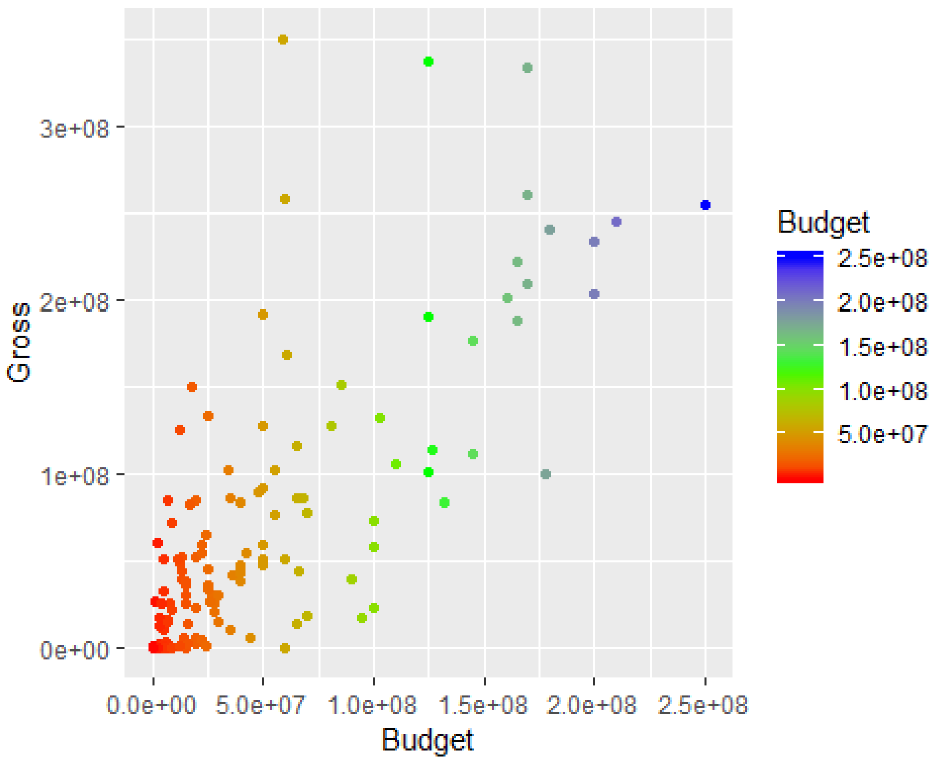


```
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014boxcox1, which=c(1,2))
```

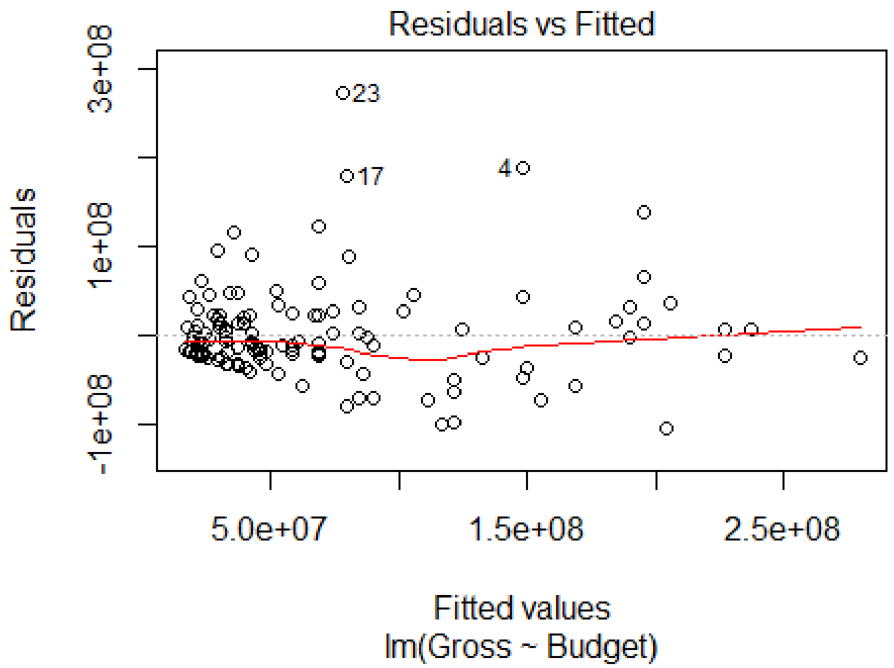


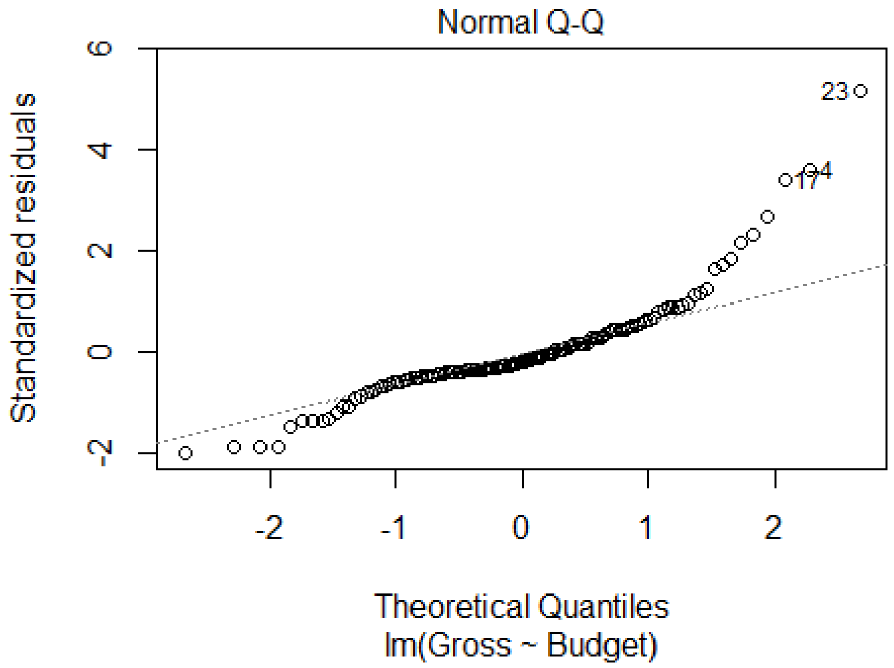


```
# Scatterplot and Regression of Gross Revenue vs. Budget
ggplot(CSM2014naomit, aes(x=Budget, y=Gross, color=Budget))+geom_point()+scale_color_gradientn(colours=rainbow(3))
```

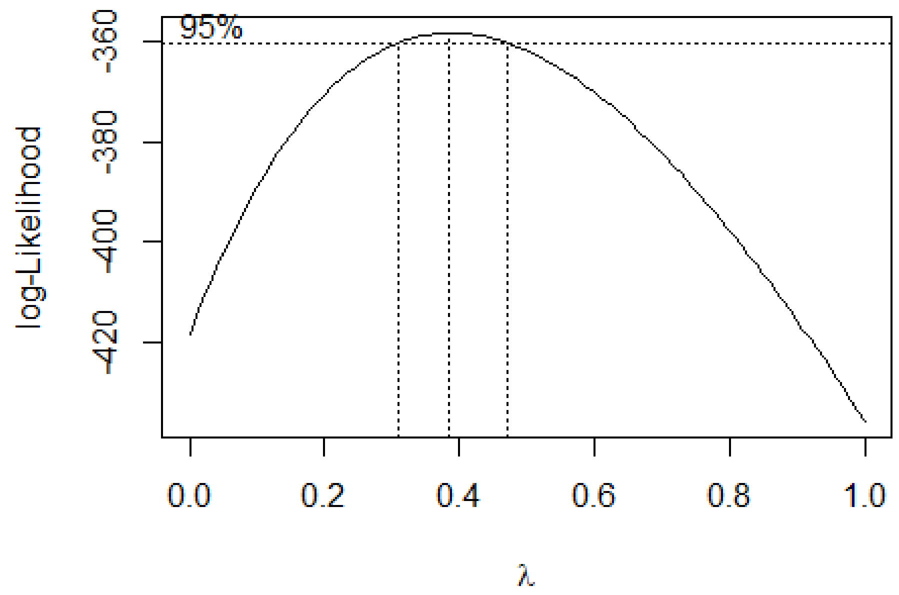


```r
CSM2014LoBF2 = lm(Gross~Budget)
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014LoBF2, which=c(1,2))
```

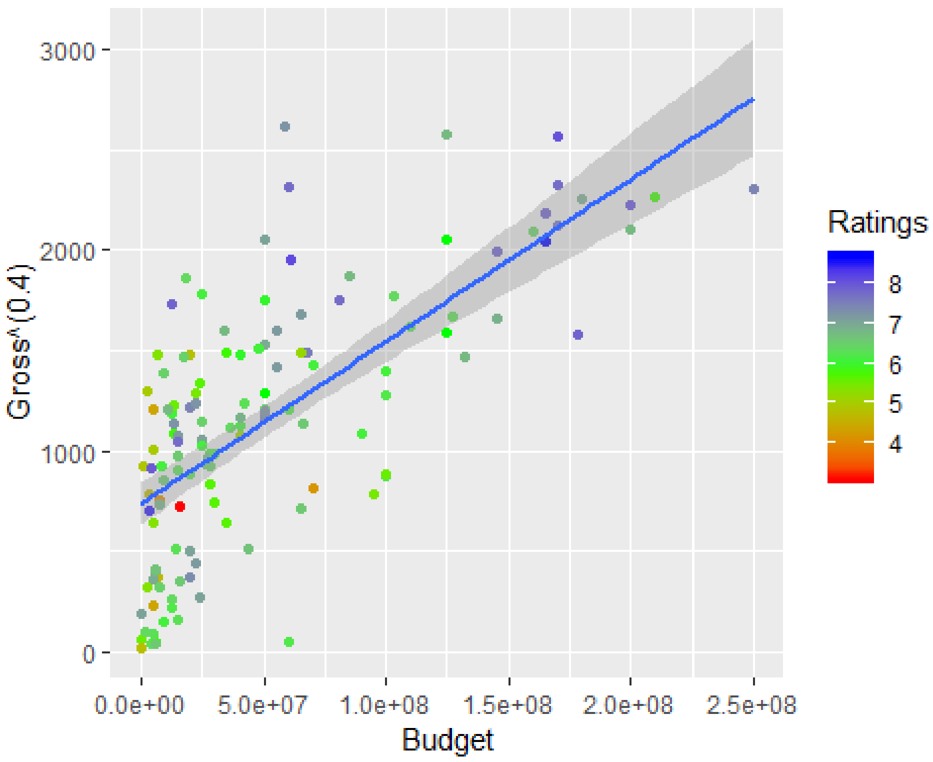




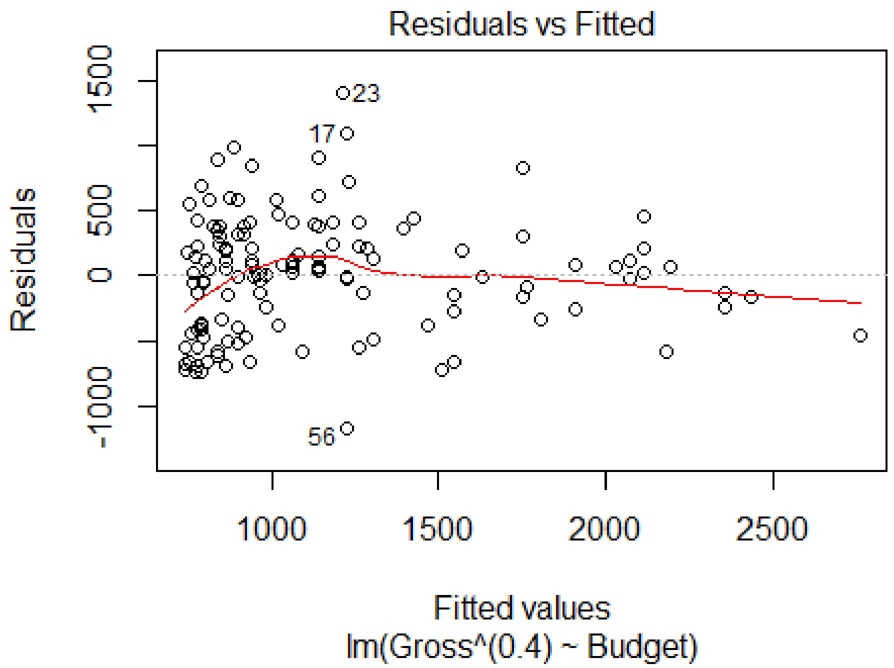
```
# Box-Cox Power Transformation to Find the Optimal Transformed Response Power
boxcox(Gross~Budget, lambda=seq(0,1,0.1))
```

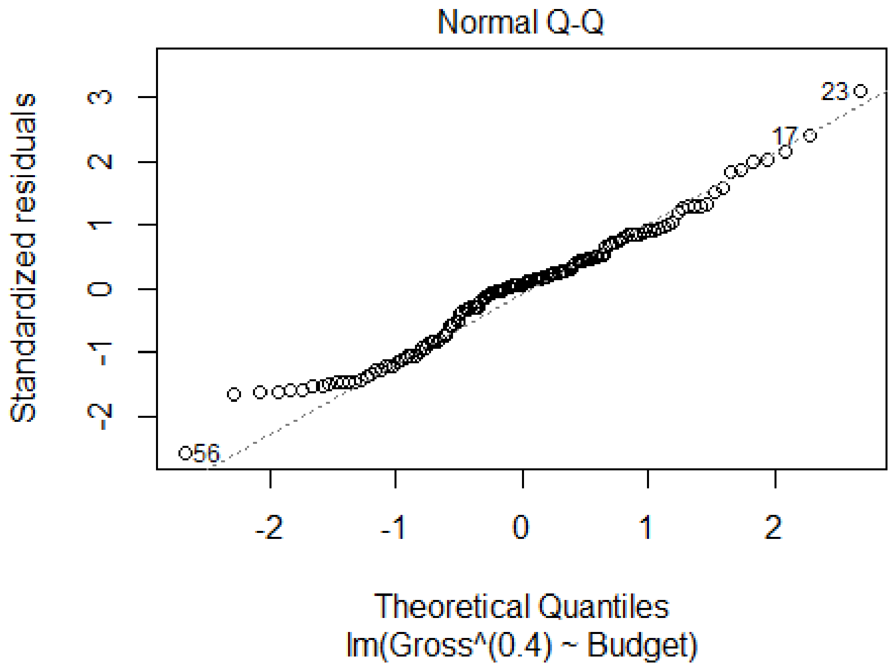


```r
# Regression with Transformed Response
CSM2014boxcox2 = lm(Gross^(0.4)~Budget)
summary(CSM2014boxcox2)

##
## Call:
## lm(formula = Gross^(0.4) ~ Budget)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1168.98  -375.28    29.41   297.49  1401.58
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.399e+02  5.403e+01   13.69   <2e-16 ***
## Budget      8.069e-06  7.201e-07   11.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 454.3 on 131 degrees of freedom
## Multiple R-squared:  0.4894, Adjusted R-squared:  0.4855
## F-statistic: 125.6 on 1 and 131 DF,  p-value: < 2.2e-16

# Scatterplot and Regression of Transformed Gross Revenue vs. Ratings
ggplot(CSM2014naomit, aes(x=Budget, y=Gross^(0.4), color=Ratings))+geom_point
()+scale_color_gradientn(colours=rainbow(3))+geom_smooth(method=lm)
```



```
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014boxcox2, which=c(1,2))
```





```
# Akaike's Information Criterion
CSM2014LoBFReduced = lm(Gross~1)
step(CSM2014LoBFReduced, scope = list(lower=CSM2014LoBFReduced, upper=CSM2014
LoBF))
```

```
## Start:  AIC=4835.15
## Gross ~ 1
##
##                          Df  Sum of Sq        RSS    AIC
## + Budget                 1 4.3960e+17 3.6559e+17 4732.1
## + Screens                1 3.1309e+17 4.9210e+17 4771.7
## + Genre                  8 2.4560e+17 5.5959e+17 4802.8
## + Ratings                1 1.3050e+17 6.7469e+17 4813.6
## + Sequel                 1 9.8037e+16 7.0715e+17 4819.9
## + Aggregate.Followers  1 7.1797e+16 7.3339e+17 4824.7
## <none>                              8.0519e+17 4835.2
## + Comments               1 1.1634e+16 7.9356e+17 4835.2
## + Likes                  1 8.8210e+15 7.9637e+17 4835.7
## + Dislikes               1 5.9750e+15 7.9922e+17 4836.2
## + Views                  1 2.5887e+15 8.0260e+17 4836.7
##
## Step:  AIC=4732.14
## Gross ~ Budget
##
##                          Df  Sum of Sq        RSS    AIC
## + Screens                1 4.5075e+16 3.2051e+17 4716.6
## + Aggregate.Followers  1 4.3029e+16 3.2256e+17 4717.5
## + Ratings                1 1.9276e+16 3.4631e+17 4726.9
## + Likes                  1 1.0522e+16 3.5506e+17 4730.3
## <none>                              3.6559e+17 4732.1
## + Comments               1 3.5346e+15 3.6205e+17 4732.8
## + Dislikes               1 1.8646e+15 3.6372e+17 4733.5
## + Views                  1 6.0294e+14 3.6498e+17 4733.9
## + Sequel                 1 2.6099e+14 3.6533e+17 4734.0
## + Genre                  8 1.5721e+16 3.4987e+17 4742.3
## - Budget                 1 4.3960e+17 8.0519e+17 4835.2
##
## Step:  AIC=4716.64
## Gross ~ Budget + Screens
##
##                          Df  Sum of Sq        RSS    AIC
## + Aggregate.Followers  1 2.8951e+16 2.9156e+17 4706.0
## + Ratings                1 2.8083e+16 2.9243e+17 4706.4
## + Likes                  1 5.4813e+15 3.1503e+17 4716.3
## <none>                              3.2051e+17 4716.6
## + Comments               1 8.6665e+14 3.1965e+17 4718.3
## + Sequel                 1 8.3708e+14 3.1967e+17 4718.3
## + Dislikes               1 3.1890e+13 3.2048e+17 4718.6
## + Views                  1 2.9287e+12 3.2051e+17 4718.6
## + Genre                  8 9.8956e+15 3.1062e+17 4728.5
## - Screens                1 4.5075e+16 3.6559e+17 4732.1
## - Budget                 1 1.7159e+17 4.9210e+17 4771.7
##
## Step:  AIC=4706.05
## Gross ~ Budget + Screens + Aggregate.Followers
```

```
##
##                        Df  Sum of Sq        RSS    AIC
## + Ratings               1 2.4420e+16 2.6714e+17 4696.4
## + Likes                 1 4.3632e+15 2.8720e+17 4706.0
## <none>                               2.9156e+17 4706.0
## + Comments              1 9.8445e+14 2.9058e+17 4707.6
## + Views                 1 6.4656e+14 2.9091e+17 4707.8
## + Dislikes              1 4.2327e+14 2.9114e+17 4707.9
## + Sequel                1 2.2998e+12 2.9156e+17 4708.0
## - Aggregate.Followers  1 2.8951e+16 3.2051e+17 4716.6
## - Screens               1 3.0996e+16 3.2256e+17 4717.5
## + Genre                 8 6.5247e+15 2.8504e+17 4719.0
## - Budget                1 1.7650e+17 4.6806e+17 4767.0
##
## Step:  AIC=4696.42
## Gross ~ Budget + Screens + Aggregate.Followers + Ratings
##
##                        Df  Sum of Sq        RSS    AIC
## <none>                               2.6714e+17 4696.4
## + Views                 1 1.7842e+15 2.6536e+17 4697.5
## + Likes                 1 1.3085e+15 2.6583e+17 4697.8
## + Sequel                1 4.7616e+14 2.6666e+17 4698.2
## + Comments              1 1.9224e+14 2.6695e+17 4698.3
## + Dislikes              1 2.3226e+13 2.6712e+17 4698.4
## - Ratings               1 2.4420e+16 2.9156e+17 4706.0
## - Aggregate.Followers  1 2.5288e+16 2.9243e+17 4706.4
## + Genre                 8 1.1456e+16 2.5568e+17 4706.6
## - Screens               1 3.8471e+16 3.0561e+17 4712.3
## - Budget                1 1.1312e+17 3.8026e+17 4741.4
```

```
##
## Call:
## lm(formula = Gross ~ Budget + Screens + Aggregate.Followers +
##     Ratings)
##
## Coefficients:
##         (Intercept)               Budget              Screens
##          -1.037e+08            7.039e-01            1.462e+04
## Aggregate.Followers              Ratings
##           2.561e+00            1.531e+07
```

```
# Bayesian Information Criterion
CSM2014regsubsets = summary(regsubsets(cbind(Ratings, Genre, Budget, Screens,
Sequel, Views, Likes, Dislikes, Comments, Aggregate.Followers), Gross^(1/3)))
CSM2014regsubsets$which
```

```
##   (Intercept) Ratings Genre Budget Screens Sequel Views Likes Dislikes
## 1        TRUE   FALSE FALSE  FALSE    TRUE  FALSE FALSE FALSE    FALSE
## 2        TRUE   FALSE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 3        TRUE    TRUE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 4        TRUE    TRUE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 5        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 6        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE  TRUE    FALSE
## 7        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE  TRUE     TRUE
## 8        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE  TRUE  TRUE     TRUE
##   Comments Aggregate.Followers
## 1    FALSE               FALSE
## 2    FALSE               FALSE
## 3    FALSE               FALSE
## 4    FALSE                TRUE
## 5    FALSE                TRUE
## 6    FALSE                TRUE
## 7    FALSE                TRUE
## 8    FALSE                TRUE
```

```
CSM2014regsubsets$adjr2
```

```
## [1] 0.5123526 0.6140083 0.6400407 0.6591498 0.6643286 0.6676344 0.6657004
## [8] 0.6666742
```

```
# Mallow's Cp Statistic
CSM2014regsubsets$which
```

```
##   (Intercept) Ratings Genre Budget Screens Sequel Views Likes Dislikes
## 1        TRUE   FALSE FALSE  FALSE    TRUE  FALSE FALSE FALSE    FALSE
## 2        TRUE   FALSE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 3        TRUE    TRUE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 4        TRUE    TRUE FALSE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 5        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE FALSE    FALSE
## 6        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE  TRUE    FALSE
## 7        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE FALSE  TRUE     TRUE
## 8        TRUE    TRUE  TRUE   TRUE    TRUE  FALSE  TRUE  TRUE     TRUE
##   Comments Aggregate.Followers
## 1    FALSE               FALSE
## 2    FALSE               FALSE
## 3    FALSE               FALSE
## 4    FALSE                TRUE
## 5    FALSE                TRUE
## 6    FALSE                TRUE
## 7    FALSE                TRUE
## 8    FALSE                TRUE
```

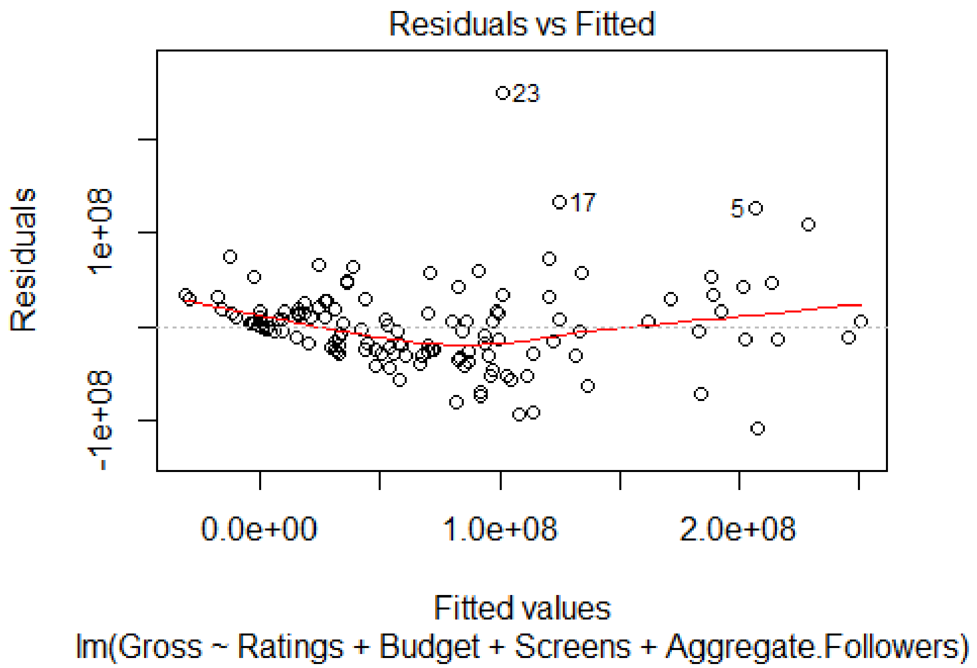
```
CSM2014regsubsets$cp
```

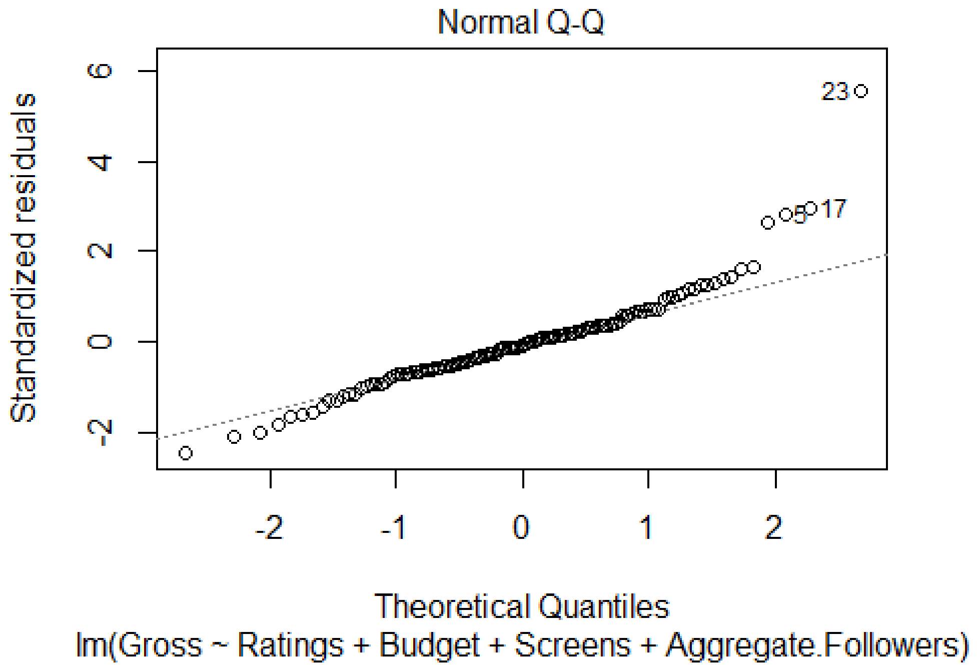
```
## [1] 61.944670 22.986315 13.794864  7.407894  6.423181  6.174809  7.903955
## [8]  8.543812
```

```
# Choose the AIC Model
CSM2014LoBFAIC = lm(Gross~Ratings+Budget+Screens+Aggregate.Followers)
summary(CSM2014LoBFAIC)
```

```
##
## Call:
## lm(formula = Gross ~ Ratings + Budget + Screens + Aggregate.Followers)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -107921021  -25614437   -2461732   17264221  248526133
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.037e+08  2.878e+07  -3.604 0.000447 ***
## Ratings              1.531e+07  4.476e+06   3.421 0.000839 ***
## Budget               7.039e-01  9.561e-02   7.362 1.95e-11 ***
## Screens              1.462e+04  3.405e+03   4.293 3.45e-05 ***
## Aggregate.Followers  2.561e+00  7.358e-01   3.481 0.000684 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45680000 on 128 degrees of freedom
## Multiple R-squared:  0.6682, Adjusted R-squared:  0.6579
## F-statistic: 64.45 on 4 and 128 DF,  p-value: < 2.2e-16
```

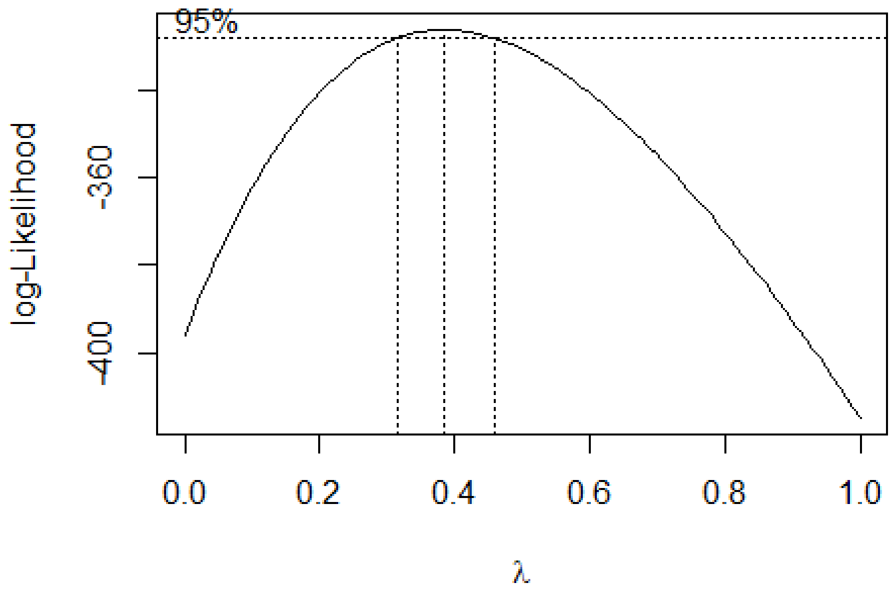
```r
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014LoBFAIC, which=c(1,2))
```





```r
# Logarithmic Transformation on Response and Predictors
CSM2014log = lm(log(Gross)~log(Ratings)+log(Budget)+log(Screens)+log(Aggregat
e.Followers))
summary(CSM2014log)$r.squared
```

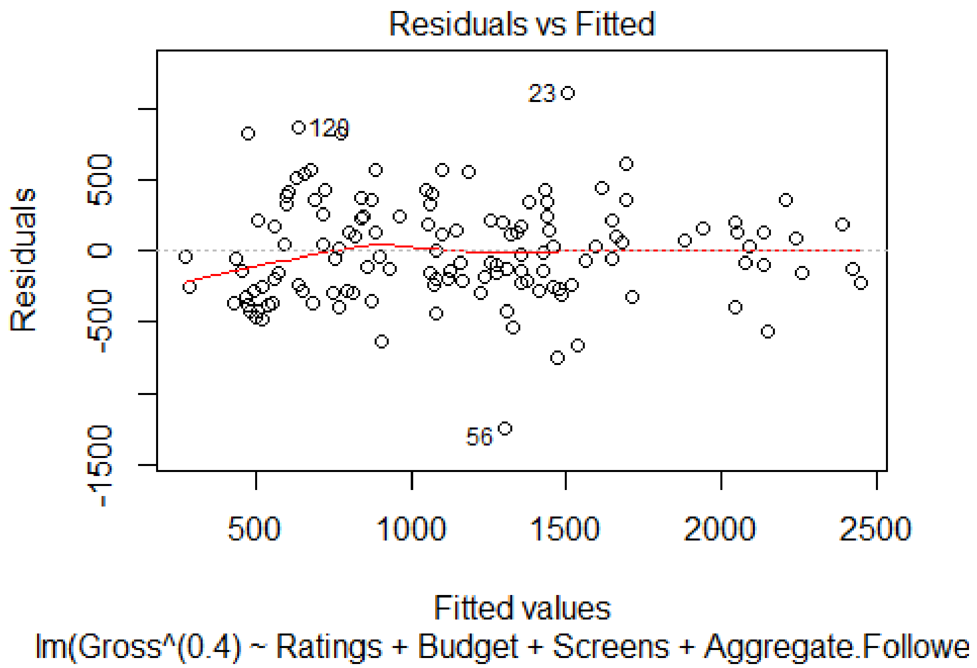
```
## [1] 0.5848727

# Box-Cox Power Transformation to Find the Optimal Transformed Response Power
boxcox(Gross~Ratings+Budget+Screens+Aggregate.Followers, lambda=seq(0,1,0.1))
```

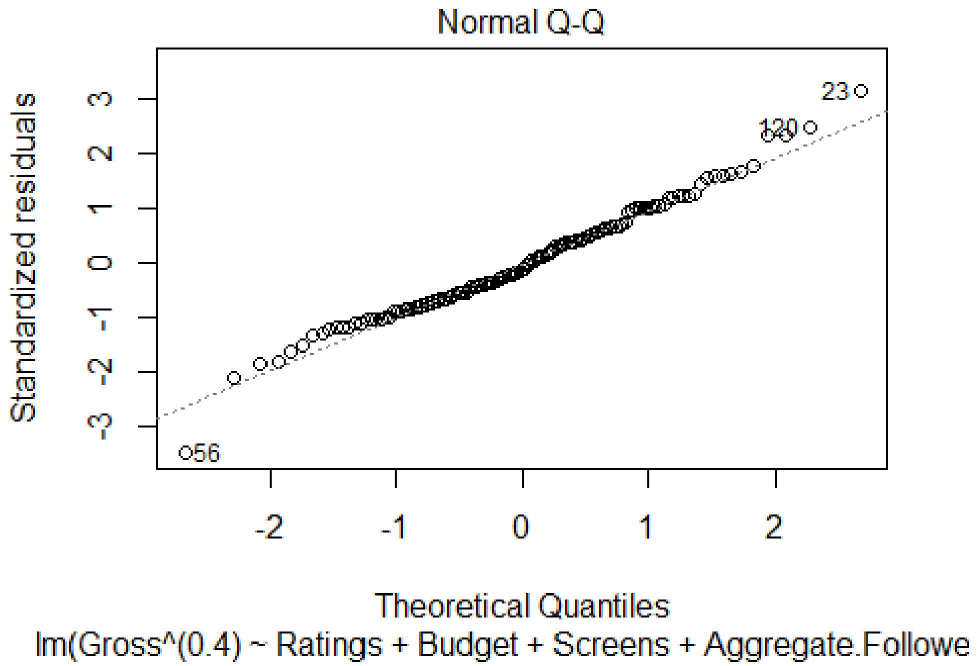


```
# Regression with Transformed Response
CSM2014boxcox = lm(Gross^(0.4)~Ratings+Budget+Screens+Aggregate.Followers)
summary(CSM2014boxcox)

## 
## Call:
## lm(formula = Gross^(0.4) ~ Ratings + Budget + Screens + Aggregate.Follower
s)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1243.75  -246.81   -36.65   221.20  1112.47 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         -2.738e+02  2.272e+02  -1.205  0.23049    
## Ratings              1.153e+02  3.535e+01   3.263  0.00141 ** 
## Budget               4.137e-06  7.550e-07   5.480 2.18e-07 ***
## Screens              1.946e-01  2.689e-02   7.236 3.77e-11 ***
## Aggregate.Followers  1.794e-05  5.810e-06   3.088  0.00247 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 360.8 on 128 degrees of freedom
## Multiple R-squared:  0.6853, Adjusted R-squared:  0.6755
## F-statistic: 69.69 on 4 and 128 DF,  p-value: < 2.2e-16
```

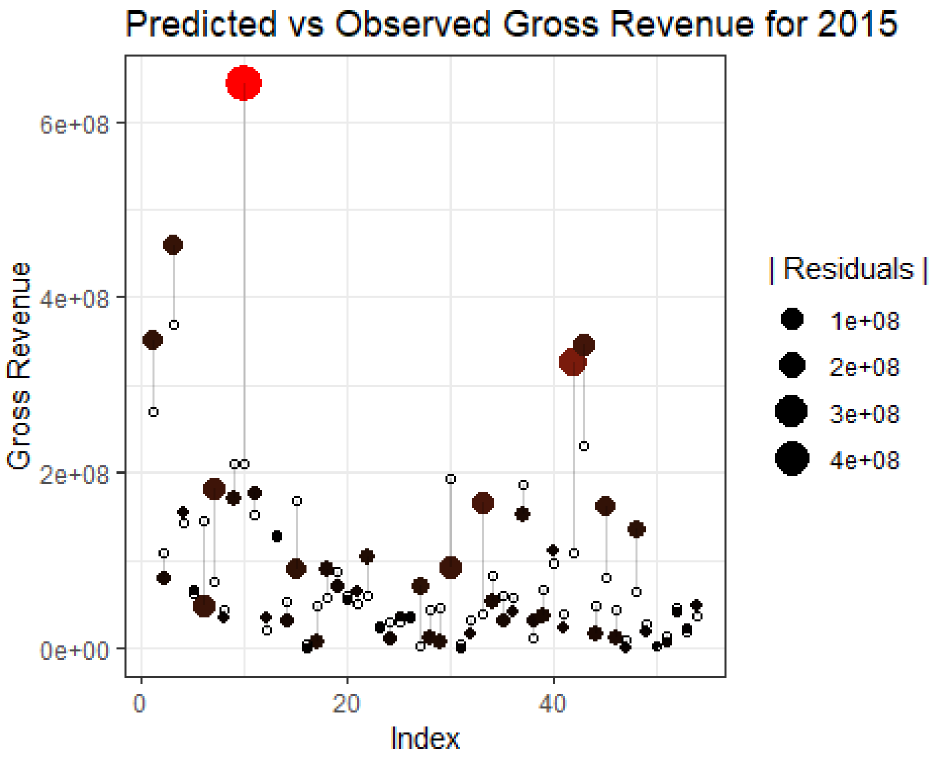
```r
# Residuals vs. Fitted and Normal Q-Q Plots
plot(CSM2014boxcox, which=c(1,2))
```





```
# Omitting Data with N/A Values in 2015 Data Set
CSM2015naomit = na.omit(CSM2015)
# Predictions on 2015 Movies
```

```r
CSM2015Prediction = predict(CSM2014boxcox, CSM2015naomit)^2.5
write.csv(CSM2015Prediction, file = "CSM2015Prediction.csv")

# Calculating Residual Values
CSM2015Resid = CSM2015Prediction - CSM2015naomit$Gross
# Plotting Residuals
ggplot(CSM2015naomit, aes(x=seq(1, length(CSM2015naomit$Gross)), y=Gross, label=CSM2015naomit$Movie))+
geom_point(aes(color=abs(CSM2015Resid), size=abs(CSM2015Resid)))+
geom_point(aes(y=CSM2015Prediction), shape=1)+
geom_segment(aes(xend=seq(1, length(CSM2015naomit$Gross)), yend=CSM2015Prediction), alpha=0.2)+
scale_color_continuous(low="black", high="red")+
labs(title="Predicted vs Observed Gross Revenue for 2015", x="Index", y="Gross Revenue", size="| Residuals |")+
theme_bw()+
guides(color = FALSE)
```


Predicted vs Observed Gross Revenue for 2015
Gross Revenue
6e+08
4e+08
2e+08
0e+00
0
20
40
Index
| Residuals |
1e+08
2e+08
3e+08
4e+08