SLR\_MLR

## Introduction:

We are interested in finding out if we can predict a film's popularity from its initial budget, the revenue it generated, the average vote it was given online, and whether the title had the word "man" or "men" in it. Our data sources is from the Kaggle website <https://www.kaggle.com/tmdb/tmdb-movie-metadata/data>, which provides a comma-separated-value matrix of different information on five thousand movies.

## Are any explanatory variables correlated?

First, we wanted to get a sense of if and how the different variables were correlated with eachother. We used pair plots between each of the four variables to visually inspect this. Most appeared to be normally distributed and not correlated. However, budget and revenue were highly correlated (r = 0.73), which is not surprising, as a higher budget usually leads to famous actors and lots of advertising.

## Transforming the data

To make sure that the technical conditions were met, we checked the residual plots of each of the variables. The residual plot of popularity versus budget was neither constant nor linear, so we transformed both variables with a logarithmic function, which made the residuals evenly scattered around 0. To do this, we had to omit all the movies that had a budget of 0$, as these would have been undefined. Similarly, we log-transformed both popularity and revenue, omitting all values of zero.

The residuals for average vote versus popularity, and whether the title contained "Men" and popularity, were non constant, so we log-transformed only popularity in those cases. These transformations were sufficient to produce evenly scattered residual plots, so we did not include any further transformations.

## Do any of the variables interact?

We were curious if any of the variables interacted with eachother to modulate how they informed the response, Popularity. To test this, we compared the full model with variations that each had an interactive term. From these, we found that budget and revenue had a significant interaction, as did revenue and vote average. The other interactions were not significant. However, the two interactive terms were not significant together, and the more significant interaction, between revenue and budget, made the original 'revenue' term not significant. Therefore, we decided to not include the interactive terms, as they make the model difficult to interpret, and reduce the significance of the original variable itself.

## Building the model:

Because our fitted model log-transformed popularity, budget, and revenue, we omitted all data points from those vectors that were 0.

# put in summary output here / tidy

The values for the log-transformed Budget variable was 0.07389; for log-transformed revenue 0.275; for average vote, 0.335; and for men, 0.1311. All values except for "Men" were significant (had p-values less than 1e-10). A doubling in budget or revenue is consistent with a multiplicative change in the median of popularity of 1.05 or 1.21, respectively. An increase in one unit of average vote is associated with a 1.4 multiplicative change in the median of popularity.

Because the p-value for the variable "Men" was 0.18, we decided that it probably does not inform a movie's popularity. Therefore, we decided to drop it from the model.

## Evaluating the variables:

To determine how necessary all the variables are for the model's success, we decided to compare the full model versus a model that only used Revenue and Average Vote as explanatory parameters. We dropped Men because it was not significant, and since Revenue and Budget are correlated, we decided to use only one. We thought that Revenue (the amount of money a movie produces) would be a better descriptor than Budget (how much money goes into a movie).

# put in anova

We used an F-test to analyze the variance of the two models. The f-statistic was 20.5, which is much higher than 1 (p value < 1e-9). Therefore, at least one of the variables in our full model is significant. We are not sure which one, if any, is insignificant, so at this point we would report the full model to our boss. Parsimony is good, but precision is better.

The of the full model is 0.461, and the adjusted is 0.46. 46% of the variation in the popularity can be explained by the four predictor variables: revenue, budget and average vote, and "Men". The adjusted R^2 penalizes for the number of variables, but since we only have four, the decrease isn't that large.

Note: this $\R^2$ does not necessarily indicate how well the model will be at predicting future data. To make sure the $\R^2$ isn't artificially inflated by outliers or leverage points, we performed diagnostic measures to determine if there are any influential points.

The first diagnostic we did was check the residual and leverage plot to identify any influential outliers. We decided to use a Cook's distance of 0.5 as a cut-off for determining influentiality; none of the points were greater than that. However, out of curiousity, we picked a data point that was close to the cut-off point. Without that data point, the estimates of the parameters did not change. So, there aren't any influential outliers in the dataset. #put in cook's distance plot

## Final model

We performed forward- and backwards- stepwise selections, which both resulted in the same model: E(Y) = -5.287 + 0.275 ln(Revenue) + 0.074 ln(Budget) + 0.335 (Vote Average) As an example of the backwards selection, we started with the full model that included all four explanatory variables. We dropped the "Men" variable first, as its F-statistic was non significant. After dropping it, all the other variables' F-statistics were significant, so we kept them.

# drop1 here

## Coefficient of partial determinations

The variability (in the natural log of the Popularity) remaining after modelling log of the Popularity using log of the Budget and Vote, is reduced by 42.4% when we include the variable log of Revenue. The variability (in the natural log of the Popularity) remaining after modelling log of the Popularity using log of the Revenue and Vote, is reduced by 1.2% when we include the variable log of Budget. The variability (in the natural log of the Popularity) remaining after modelling log of the Popularity using log of the Budget and log of Revenue, is reduced by 9.65% when we include the variable of Vote. This gives us a sense that the variable Revenue is most important of the three variables in predicting Popularity.

## Interpreting the final model

Our final model included three explanatory variables to predict a movie's popularity: its budget, its average vote, and the revenue it made. After testing variables' significance, we decided to drop the variable "Man", a categorical measure of whether or not a film's title had "Man" or "Men" in it (eg "IronMan"; "Spider-Man"). We originally thought that movies with "Man" in their title woudld be action or superhero movies, which are usually popular and high-profile productions. Also, we were interested in whether people are biased in their movie choices and would rather see movies featuring Men.

With the three explanatory variables, our model explains 46% of the variance in a movie's popularity. ALthough this may seem low, the popularity score is based on activities like views, votes, and being added to users' watchlists, which are all human behavior and hard to predict mathematically.

## Using our model to predict popularity

Finally, we were interested in seeing how well our model performs, by using it to predict popularity.

We decided to test it on our favorite movie, the classic "Evan Almighty." Evan had a budget of $175M, generated revenue of $173M, and an average vote of 5.3. Plugging those values into our model gives us a mean prediction value of : ln (Expected Popularity) = -5.287 + (0.275) ln(175 000 000) + 0.074 ln (173 000 000 ) + 0.335 (5.3) = 3.12, with a prediction interval of 1.49 to 4.75 and a confidence interval of 3.07 to 3.18. These values can be interpreted as a fitted popularity score of 22.6, with a 95% confidence interval of 21.5 to 24, and 95% prediction interval 4.44 and 116. We are 95% confident that the true mean popularity score lies between 21.5 and 24, and 95% of the popularity scores of movies with the given revenue, budget and vote average, are between 4.44 and 116. The actual popularity score of Evan Almighty was 27.02, which falls within the prediction interval.

## Conclusions

Our aim for this project was to predict a film's popularity given its revenue, budget, average vote, and whether it had the word "Man" in its title. We were able to build a model using the first three variables that successfully explains 46% of the variance within the popularity scores of a database of 5,000 movies. There are probably several factors that describe a movie's popularity, including ones that were not in the dataset itself (i.e. number of awards) and ones that were too complex to code into a model (i.e. genre). However, despite these additional variables that would probably have contributed to a model's accuracy, we were able to reasonably predict a movie's popularity within a 95% prediction interval.

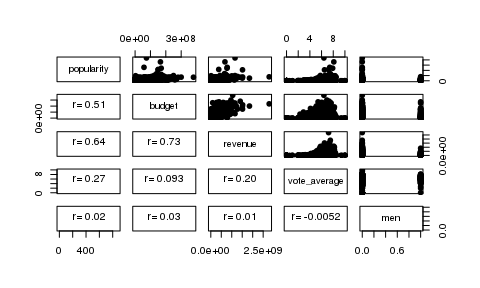
## Addendum: Multiple comparisons

We used a total of 13 hypothesis tests during the course of the experiment. When we adjust our p-values by multiplying by this number, 13, none reach significance. Therefore, although we tested for different hypotheses at the same time, we do not run into the problem of multiple comparisons.

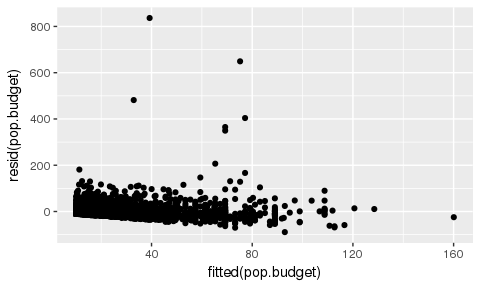
movies <- read\_csv("~/LM HW/SLM\_Project/tmdb\_5000\_movies-OG.csv")  
  
#Add variable if "man" or "men" is in the title  
library(stringr)  
test <- c("superman", "man", "table", "spider-man", "supermen","men")  
grepl("man|men", test)

## [1] TRUE TRUE FALSE TRUE TRUE TRUE

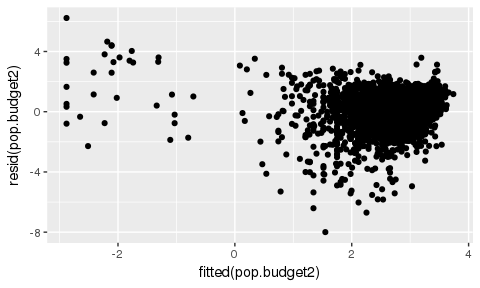
is.man <- grepl("man|men", movies$title)  
men.true <- c()  
for(i in is.man){  
 if(i == TRUE){  
 men.true <- append(1, men.true)  
 }else if (i == FALSE){  
 men.true <- append(0, men.true)  
 } else{  
 men.true <- append(3, men.true)  
 }  
}  
movies$men = rev(men.true)  
  
#Pairs plot  
panel.cor <- function(x, y, digits = 2, cex.cor, ...)  
{  
 usr <- par("usr"); on.exit(par(usr))  
 par(usr = c(0, 1, 0, 1))  
 # correlation coefficient  
 r <- cor(x, y)  
 txt <- format(c(r, 0.123456789), digits = digits)[1]  
 txt <- paste("r= ", txt, sep = "")  
 text(0.5, 0.6, txt)  
  
}  
pairs(~popularity + budget + revenue + vote\_average + men, data=movies, lower.panel=panel.cor, pch=19)



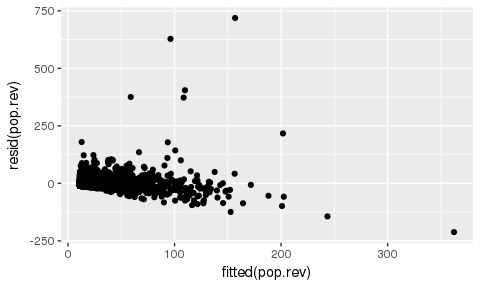
#residual plot for popularity vs. budget  
pop.budget <- lm(popularity ~ budget, data = movies)  
ggplot(pop.budget, aes(x=fitted(pop.budget), y=resid(pop.budget))) + geom\_point()



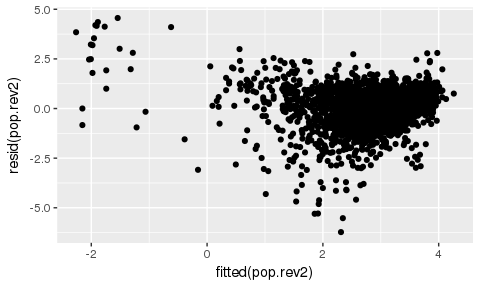
#requires log transformation so removing 0 values  
budget <- movies %>%  
 select(popularity, budget) %>%  
 filter(budget > 0, popularity > 0)  
  
pop.budget2 <- lm(log(popularity) ~ log(budget), data = budget)  
ggplot(pop.budget2, aes(x=fitted(pop.budget2), y=resid(pop.budget2))) + geom\_point()



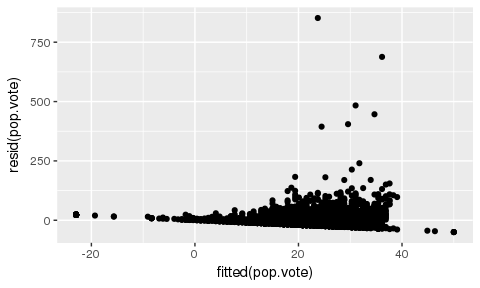
#residual plot for popularity vs. revenue  
pop.rev <- lm(popularity ~ revenue, data = movies)  
ggplot(pop.rev, aes(x=fitted(pop.rev), y=resid(pop.rev))) + geom\_point()



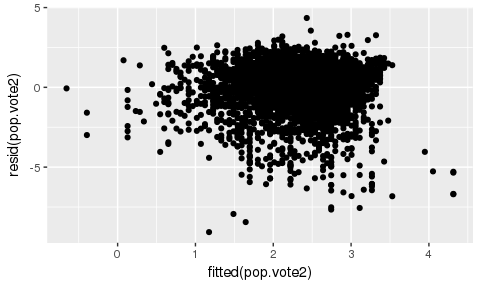
#remove 0 values and log transform popularity and revenue  
revenue <- movies %>%  
 select(popularity, revenue) %>%  
 filter(revenue > 0, popularity > 0)  
pop.rev2 <- lm(log(popularity) ~ log(revenue), data = revenue)  
ggplot(pop.rev2, aes(x=fitted(pop.rev2), y=resid(pop.rev2))) + geom\_point()



#residual plot for popularity vs. vote average  
pop.vote <- lm(popularity ~ vote\_average, data = movies)  
ggplot(pop.vote, aes(x=fitted(pop.vote), y=resid(pop.vote))) + geom\_point()



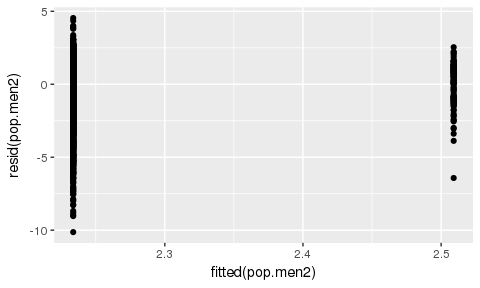
vote <- movies %>%  
 select(popularity, vote\_average) %>%  
 filter(popularity > 0, vote\_average >0)  
pop.vote2 <- lm(log(popularity) ~ vote\_average, data = vote)  
ggplot(pop.vote2, aes(x=fitted(pop.vote2), y=resid(pop.vote2))) + geom\_point()



#residual plot for popularity and men  
pop.men <- lm(popularity ~ men, data = movies)  
ggplot(pop.men, aes(x=fitted(pop.men), y=resid(pop.men))) + geom\_point()



men <- movies %>%  
 select(popularity, men) %>%  
 filter(popularity > 0)  
pop.men2 <- lm(log(popularity) ~ men, data = men)  
ggplot(pop.men2, aes(x=fitted(pop.men2), y=resid(pop.men2))) + geom\_point()



#Create LM  
movies.new <- movies %>%  
 select(popularity, men, budget, revenue, vote\_average) %>%  
 filter(popularity > 0, budget > 0, revenue > 0)  
movies.lm <- lm(log(popularity) ~ log(budget) + log(revenue) + vote\_average + men, data=movies.new)  
summary(movies.lm)

##   
## Call:  
## lm(formula = log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men, data = movies.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.092 -0.343 0.087 0.452 4.688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.28373 0.18705 -28.25 < 2e-16 \*\*\*  
## log(budget) 0.07389 0.01181 6.26 4.4e-10 \*\*\*  
## log(revenue) 0.27523 0.00951 28.93 < 2e-16 \*\*\*  
## vote\_average 0.33575 0.01716 19.56 < 2e-16 \*\*\*  
## men 0.13110 0.09833 1.33 0.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.813 on 3224 degrees of freedom  
## Multiple R-squared: 0.461, Adjusted R-squared: 0.46   
## F-statistic: 690 on 4 and 3224 DF, p-value: <2e-16

#interaction  
movies.br.int <- lm(log(popularity) ~ log(budget) \* log(revenue) + vote\_average + men, data=movies.new)  
movies.rv.int <- lm(log(popularity) ~ log(budget) + log(revenue) \* vote\_average + men, data=movies.new)  
movies.bv.int <- lm(log(popularity) ~ log(budget) \* vote\_average + log(revenue) + men, data=movies.new)  
movies.int <- lm(log(popularity) ~ log(budget) \* vote\_average + log(revenue)\*log(budget) + men, data=movies.new)  
anova(movies.lm, movies.br.int)

## Analysis of Variance Table  
##   
## Model 1: log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men  
## Model 2: log(popularity) ~ log(budget) \* log(revenue) + vote\_average +   
## men  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 3224 2132   
## 2 3223 2008 1 125 201 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.lm, movies.rv.int)

## Analysis of Variance Table  
##   
## Model 1: log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men  
## Model 2: log(popularity) ~ log(budget) + log(revenue) \* vote\_average +   
## men  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 3224 2132   
## 2 3223 2123 1 9.79 14.9 0.00012 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.lm, movies.bv.int)

## Analysis of Variance Table  
##   
## Model 1: log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men  
## Model 2: log(popularity) ~ log(budget) \* vote\_average + log(revenue) +   
## men  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 3224 2132   
## 2 3223 2132 1 0.106 0.16 0.69

anova(movies.int, movies.lm)

## Analysis of Variance Table  
##   
## Model 1: log(popularity) ~ log(budget) \* vote\_average + log(revenue) \*   
## log(budget) + men  
## Model 2: log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 3222 2004   
## 2 3224 2132 -2 -128 103 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.br.int)

## Analysis of Variance Table  
##   
## Response: log(popularity)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(budget) 1 716 716 1149.85 <2e-16 \*\*\*  
## log(revenue) 1 855 855 1372.45 <2e-16 \*\*\*  
## vote\_average 1 253 253 405.85 <2e-16 \*\*\*  
## men 1 1 1 1.89 0.17   
## log(budget):log(revenue) 1 125 125 200.54 <2e-16 \*\*\*  
## Residuals 3223 2008 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.rv.int)

## Analysis of Variance Table  
##   
## Response: log(popularity)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(budget) 1 716 716 1087.49 < 2e-16 \*\*\*  
## log(revenue) 1 855 855 1298.02 < 2e-16 \*\*\*  
## vote\_average 1 253 253 383.84 < 2e-16 \*\*\*  
## men 1 1 1 1.79 0.18161   
## log(revenue):vote\_average 1 10 10 14.87 0.00012 \*\*\*  
## Residuals 3223 2123 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.bv.int)

## Analysis of Variance Table  
##   
## Response: log(popularity)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(budget) 1 716 716 1082.55 <2e-16 \*\*\*  
## vote\_average 1 553 553 835.63 <2e-16 \*\*\*  
## log(revenue) 1 555 555 838.59 <2e-16 \*\*\*  
## men 1 1 1 1.78 0.18   
## log(budget):vote\_average 1 0 0 0.16 0.69   
## Residuals 3223 2132 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(movies.int)

## Analysis of Variance Table  
##   
## Response: log(popularity)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(budget) 1 716 716 1151.37 <2e-16 \*\*\*  
## vote\_average 1 553 553 888.75 <2e-16 \*\*\*  
## log(revenue) 1 555 555 891.90 <2e-16 \*\*\*  
## men 1 1 1 1.89 0.17   
## log(budget):vote\_average 1 0 0 0.17 0.68   
## log(budget):log(revenue) 1 128 128 205.88 <2e-16 \*\*\*  
## Residuals 3222 2004 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(movies.int)

##   
## Call:  
## lm(formula = log(popularity) ~ log(budget) \* vote\_average + log(revenue) \*   
## log(budget) + men, data = movies.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.008 -0.340 0.056 0.441 4.277   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.03478 1.00732 -3.01 0.0026 \*\*   
## log(budget) -0.08341 0.06130 -1.36 0.1737   
## vote\_average 0.68564 0.15358 4.46 8.3e-06 \*\*\*  
## log(revenue) -0.04292 0.02401 -1.79 0.0739 .   
## men 0.11438 0.09537 1.20 0.2305   
## log(budget):vote\_average -0.02107 0.00919 -2.29 0.0220 \*   
## log(budget):log(revenue) 0.02020 0.00141 14.35 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.789 on 3222 degrees of freedom  
## Multiple R-squared: 0.494, Adjusted R-squared: 0.493   
## F-statistic: 523 on 6 and 3222 DF, p-value: <2e-16

#betas  
summary(movies.br.int)

##   
## Call:  
## lm(formula = log(popularity) ~ log(budget) \* log(revenue) + vote\_average +   
## men, data = movies.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.981 -0.339 0.056 0.442 4.267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.8793 0.3601 -2.44 0.015 \*   
## log(budget) -0.2133 0.0233 -9.16 <2e-16 \*\*\*  
## log(revenue) -0.0355 0.0238 -1.49 0.136   
## vote\_average 0.3358 0.0167 20.17 <2e-16 \*\*\*  
## men 0.1173 0.0954 1.23 0.219   
## log(budget):log(revenue) 0.0198 0.0014 14.16 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.789 on 3223 degrees of freedom  
## Multiple R-squared: 0.493, Adjusted R-squared: 0.492   
## F-statistic: 626 on 5 and 3223 DF, p-value: <2e-16

tidy(movies.br.int)

## term estimate std.error statistic p.value  
## 1 (Intercept) -0.8793 0.3601 -2.44 1.47e-02  
## 2 log(budget) -0.2133 0.0233 -9.16 9.14e-20  
## 3 log(revenue) -0.0355 0.0238 -1.49 1.36e-01  
## 4 vote\_average 0.3358 0.0167 20.17 2.83e-85  
## 5 men 0.1173 0.0954 1.23 2.19e-01  
## 6 log(budget):log(revenue) 0.0198 0.0014 14.16 3.29e-44

#LM minus budget and men   
movies.lm.red <- lm(log(popularity) ~ log(revenue) + vote\_average, data=movies.new)  
  
#F test comparing full and reduced model without budget and men  
anova(movies.lm, movies.lm.red)

## Analysis of Variance Table  
##   
## Model 1: log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men  
## Model 2: log(popularity) ~ log(revenue) + vote\_average  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 3224 2132   
## 2 3226 2160 -2 -27.1 20.5 1.4e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

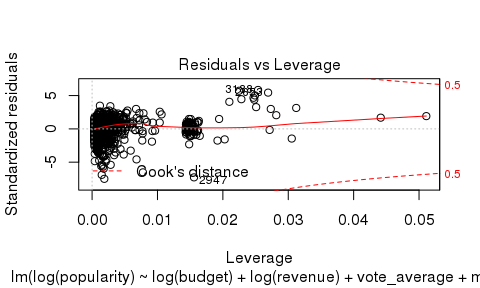
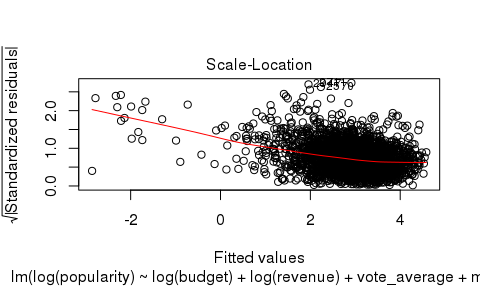
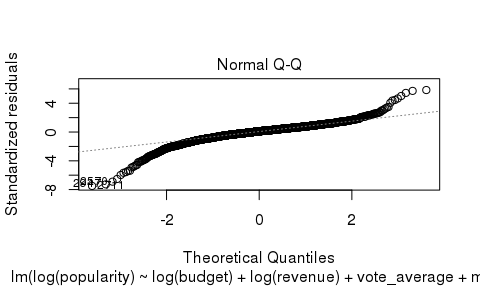
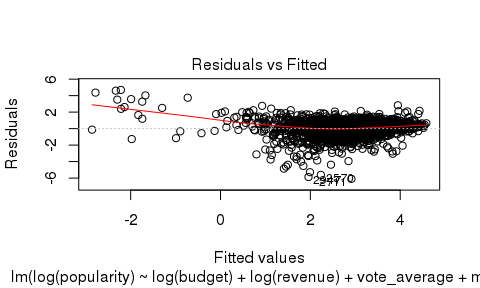
#R2 and adj R2  
summary(movies.lm)

##   
## Call:  
## lm(formula = log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men, data = movies.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.092 -0.343 0.087 0.452 4.688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.28373 0.18705 -28.25 < 2e-16 \*\*\*  
## log(budget) 0.07389 0.01181 6.26 4.4e-10 \*\*\*  
## log(revenue) 0.27523 0.00951 28.93 < 2e-16 \*\*\*  
## vote\_average 0.33575 0.01716 19.56 < 2e-16 \*\*\*  
## men 0.13110 0.09833 1.33 0.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.813 on 3224 degrees of freedom  
## Multiple R-squared: 0.461, Adjusted R-squared: 0.46   
## F-statistic: 690 on 4 and 3224 DF, p-value: <2e-16

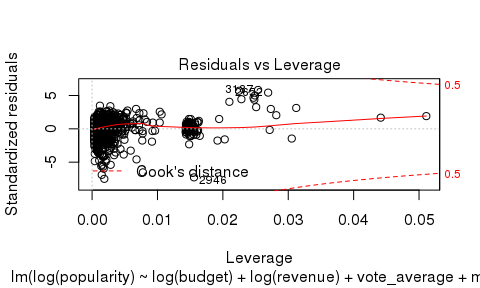
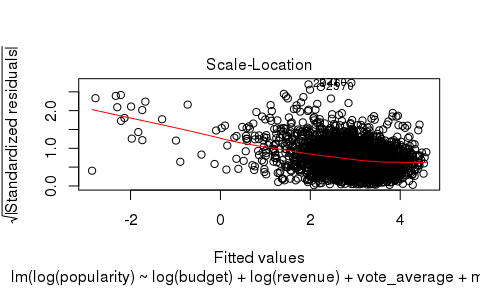
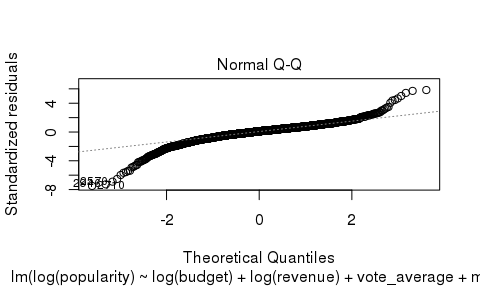
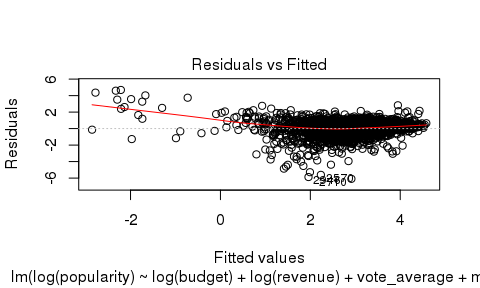
#remove influential point  
movies.new2 <- movies.new[-c(2652),]  
movies.lm2 <- lm(log(popularity) ~ log(budget) + log(revenue) + vote\_average + men, data=movies.new2)  
summary(movies.lm2)

##   
## Call:  
## lm(formula = log(popularity) ~ log(budget) + log(revenue) + vote\_average +   
## men, data = movies.new2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.092 -0.343 0.087 0.452 4.685   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.27857 0.18706 -28.22 < 2e-16 \*\*\*  
## log(budget) 0.07379 0.01180 6.25 4.6e-10 \*\*\*  
## log(revenue) 0.27516 0.00951 28.93 < 2e-16 \*\*\*  
## vote\_average 0.33543 0.01716 19.55 < 2e-16 \*\*\*  
## men 0.13077 0.09831 1.33 0.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.813 on 3223 degrees of freedom  
## Multiple R-squared: 0.461, Adjusted R-squared: 0.46   
## F-statistic: 689 on 4 and 3223 DF, p-value: <2e-16

#residual plots  
plot(movies.lm)



plot(movies.lm2)



#add   
add1(lm(log(popularity)~log(revenue), data=movies.new), log(popularity) ~ log(revenue) + vote\_average + log(budget) + men, test="F")

## Single term additions  
##   
## Model:  
## log(popularity) ~ log(revenue)  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 2387 -971   
## vote\_average 1 227.8 2160 -1293 340.24 <2e-16 \*\*\*  
## log(budget) 1 0.9 2386 -970 1.25 0.26   
## men 1 0.8 2387 -970 1.09 0.30   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

add1(lm(log(popularity)~log(revenue)+vote\_average, data=movies.new), log(popularity) ~ log(revenue) + vote\_average + log(budget) + men, test="F")

## Single term additions  
##   
## Model:  
## log(popularity) ~ log(revenue) + vote\_average  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 2160 -1293   
## log(budget) 1 25.96 2134 -1330 39.23 4.3e-10 \*\*\*  
## men 1 1.23 2158 -1293 1.83 0.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

add1(lm(log(popularity)~log(revenue)+vote\_average+log(budget), data=movies.new), log(popularity) ~ log(revenue) + vote\_average + log(budget) + men, test="F")

## Single term additions  
##   
## Model:  
## log(popularity) ~ log(revenue) + vote\_average + log(budget)  
## Df Sum of Sq RSS AIC F value Pr(>F)  
## <none> 2134 -1330   
## men 1 1.18 2132 -1330 1.78 0.18

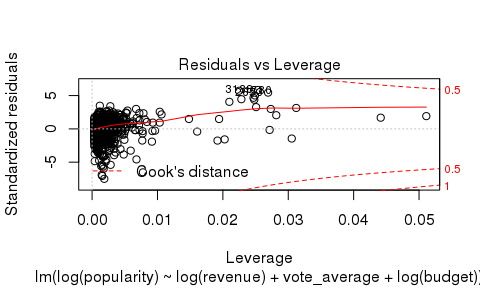
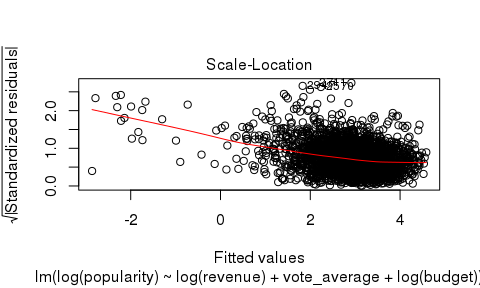
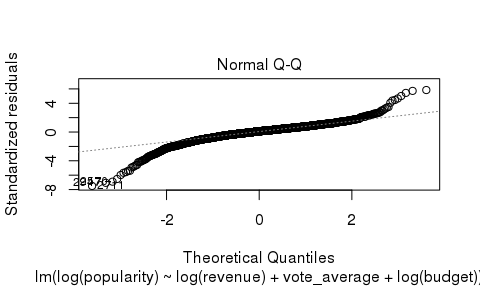
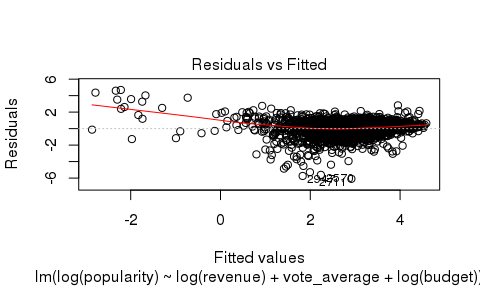
#drop   
drop1(lm(log(popularity) ~ log(revenue) + vote\_average + log(budget) + men, data=movies.new), test="F")

## Single term deletions  
##   
## Model:  
## log(popularity) ~ log(revenue) + vote\_average + log(budget) +   
## men  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 2132 -1330   
## log(revenue) 1 554 2686 -586 837.03 < 2e-16 \*\*\*  
## vote\_average 1 253 2386 -969 382.78 < 2e-16 \*\*\*  
## log(budget) 1 26 2158 -1293 39.16 4.4e-10 \*\*\*  
## men 1 1 2134 -1330 1.78 0.18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

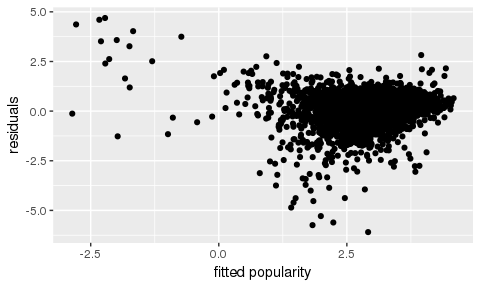
drop1(lm(log(popularity) ~ log(revenue) + vote\_average + log(budget), data=movies.new), test="F")

## Single term deletions  
##   
## Model:  
## log(popularity) ~ log(revenue) + vote\_average + log(budget)  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 2134 -1330   
## log(revenue) 1 555 2688 -586 838.6 < 2e-16 \*\*\*  
## vote\_average 1 253 2386 -970 382.1 < 2e-16 \*\*\*  
## log(budget) 1 26 2160 -1293 39.2 4.3e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

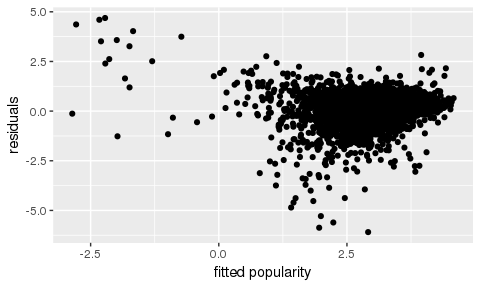
final.lm <- lm(log(popularity) ~ log(revenue) + vote\_average + log(budget), data=movies.new)  
plot(final.lm)



ggplot(final.lm, aes(x=fitted(final.lm), y=resid(final.lm))) + geom\_point() + xlab("fitted popularity") + ylab("residuals")



ggplot(movies.lm, aes(x=fitted(movies.lm), y=resid(movies.lm))) + geom\_point() + xlab("fitted popularity") + ylab("residuals")



#final model  
summary(final.lm)

##   
## Call:  
## lm(formula = log(popularity) ~ log(revenue) + vote\_average +   
## log(budget), data = movies.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.094 -0.346 0.087 0.450 4.690   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.28458 0.18708 -28.25 < 2e-16 \*\*\*  
## log(revenue) 0.27548 0.00951 28.96 < 2e-16 \*\*\*  
## vote\_average 0.33547 0.01716 19.55 < 2e-16 \*\*\*  
## log(budget) 0.07396 0.01181 6.26 4.3e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.813 on 3225 degrees of freedom  
## Multiple R-squared: 0.461, Adjusted R-squared: 0.46   
## F-statistic: 919 on 3 and 3225 DF, p-value: <2e-16

newmovie <- data.frame(budget=c(175000000), revenue=c(173000000), vote\_average=c(5.3))  
crit\_val <- qt(.975, glance(final.lm)$df.resid)  
movie\_pred <- augment(final.lm, newdata=newmovie, type.predict = "response")  
# the SE of the predictions also include the overall variability of the model  
.se.pred <- sqrt(glance(final.lm)$sigma^2 + movie\_pred$.se.fit)  
movie\_pred <- movie\_pred %>%  
mutate(lower\_PI = .fitted - crit\_val \* .se.pred,  
upper\_PI = .fitted + crit\_val \* .se.pred,  
lower\_CI = .fitted - crit\_val \* .se.fit,  
upper\_CI = .fitted + crit\_val \* .se.fit)  
movie\_pred

## budget revenue vote\_average .fitted .se.fit lower\_PI upper\_PI  
## 1 1.75e+08 1.73e+08 5.3 3.12 0.0278 1.49 4.75  
## lower\_CI upper\_CI  
## 1 3.07 3.18

#coefficients of partial determination  
anova.lm <- anova(final.lm)  
anova.lm

## Analysis of Variance Table  
##   
## Response: log(popularity)  
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(revenue) 1 1570 1570 2373.3 < 2e-16 \*\*\*  
## vote\_average 1 228 228 344.3 < 2e-16 \*\*\*  
## log(budget) 1 26 26 39.2 4.3e-10 \*\*\*  
## Residuals 3225 2134 1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova.lm[1,2]/(anova.lm[1,2]+anova.lm[4,2])

## [1] 0.424

anova.lm[3,2]/(anova.lm[3,2]+anova.lm[4,2])

## [1] 0.012

anova.lm[2,2]/(anova.lm[2,2]+anova.lm[4,2])

## [1] 0.0965