**People Love Drama:**

**A Comparison of Average IMDB User Ratings Across Movie Genres**

**Introduction**

The Internet Movie Database (IMDB) collects a wide range of data on movies to include user ratings of the movies. The website offers a lot of ways to search and sort by different data points for the movie, which allows users to drill down to the top rated movie by genre and sub-genre or by year. However, it does not offer any comparisons between genres. Therefore, the primary question of interest in this study is whether there is any difference among the mean user ratings of movies in three genres: drama, action, and comedy. As secondary concerns, the impacts of year, gross domestic ticket sales and profitability were tested as covariates for average user rating in case there was an association that impacted mean user rating.

For the genre main effect, the null hypothesis is that the mean user ratings for all three genres are the same:

The alternative hypothesis is that at least one of three genre’s mean user rating is different:

**Methods**

This study was conducted by using IMDB’s “Random Title” feature on its website to select at random[[1]](#footnote-1) the first 20 movies that fell into one of the three genres for a total of 60 movies from the almost 350,000 feature length movies in the database as of April 2016.[[2]](#footnote-2) The following information was extracted from IMDB for each movie: movie title (Movie), year released (Year), genre (Genre), average IMDB user rating (Rating), estimated budget in US dollars (EstBudgetUSD), and gross domestic ticket sales in US dollars (grossUSD). Subtracting the estimated budget from the gross domestic ticket sales generated an additional variable, profitability (profitUSD).

This experiment has one factor, genre. This factor is treated as a random effect. GrossUSD is treated as continuous covariate. The response variable is IMDB user ratings. Additional analysis was done to determine whether Year and profitUSD should be treated as covariates.

As part of an exploratory data analysis, a simple linear regression was run for Rating versus Year, grossUSD, and profitUSD for each genre separately. Both the grossUSD and profitUSD had a significant positive linear relationship for only the action genre, which means we can reject the null hypothesis that there is no linear relation with average user rating and the variables should be considered as covariates. Upon further investigation, the two variables: grossUSD and profitUSD were 83% correlated. Therefore, profitUSD was dropped as a covariate because the two variables explain essentially the same effect. The regression slopes for Year were all zero, which means we could not reject the null hypothesis that there is no linear relationship between year and average user rating. Thus the variable was not included in the final model. Graphical output for the regression and tests of significance are included in the Appendix. A significant interaction effect between genre and the grossUSD covariate revealed that the linear relationships do not have equal slopes for all genres. Therefore, an ANCOVA analysis was conducted with an unequal slopes model.

The ANCOVA test results were produced using R version 3.2.3. The R code can be found in the Appendix. The ANCOVA tests were also run in Minitab 17.2.1 for verification.

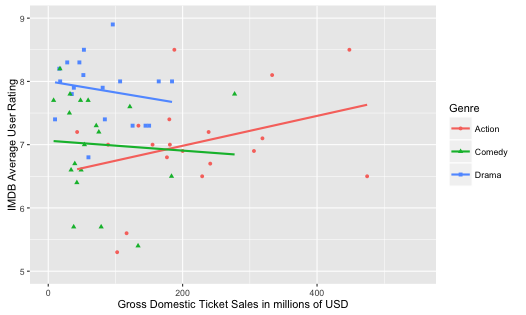
**Results**

The unequal slopes ANCOVA model produced the following regression equations for finding the average user rating for each genre:

Action

Drama

Comedy



However, the interaction term for the comedy genre was not significant. For the unequal slopes model the differences in the predicted values were compared for three different levels GrossUSD in millions of USD: 50, 150, and 250.

The differences in the predicted average user rating between drama and action at the three levels:

$50 mil $150 mil $250 mil

1.5070433 0.4082012 -0.6906409

The differences in predicted average user rating for drama and comedy at the three levels:

$50 mil $150 mil $250 mil

1.0819786 0.3217191 -0.4385405

The differences in predicted values for action and comedy at the three levels

$50 mil $150 mil $250 mil

-0.42506471 -0.08648215 0.25210041

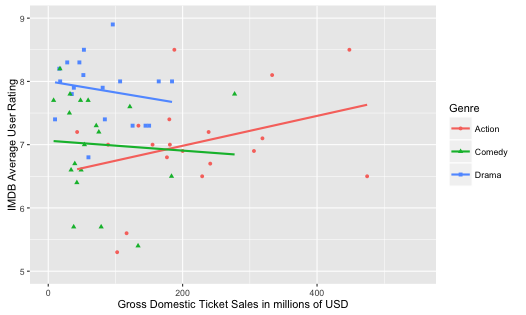
**Conclusion**

From the regression equations it appears that the average user rating for movies in the drama genre are generally higher mean user ratings than both comedy and action movies. However, as gross domestic ticket sales increase, the average user rating of movies in the action genre increases while the ratings for action and comedy movies decrease. One problem with this analysis is the variance differs among the three genres. Therefore, a comparison for the average user rating at multiple levels of GrossUSD for the entire range of action movies would include levels outside the range of the original data for comedy and action where the predictive power of the linear models is unreliable. In addition, the data set was not evaluated for outliers and their impact on the models.

**Appendix**

**Scatterplot of User Ratings vs. Gross Ticket Sales, grouped by genre:**

Not all the slopes are zero, and the slopes are not equal based on genre.



[1] "Movie Genre Action"

Call:

lm(formula = Rating ~ GrossUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-1.67024 -0.51078 0.03804 0.47706 1.53088

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.183362 0.394502 15.67 6.17e-12 \*\*\*

GrossUSD 0.004187 0.001474 2.84 0.0109 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8503 on 18 degrees of freedom

Multiple R-squared: 0.3094, Adjusted R-squared: 0.2711

F-statistic: 8.065 on 1 and 18 DF, p-value: 0.01087

[1] "Movie Genre Drama"

Call:

lm(formula = Rating ~ GrossUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-3.00977 -0.13973 0.07681 0.40351 1.31343

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.239826 0.375301 21.955 1.91e-14 \*\*\*

GrossUSD -0.006802 0.003776 -1.801 0.0884 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9224 on 18 degrees of freedom

Multiple R-squared: 0.1528, Adjusted R-squared: 0.1057

F-statistic: 3.245 on 1 and 18 DF, p-value: 0.08841

[1] "Movie Genre Comedy"

Call:

lm(formula = Rating ~ GrossUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-3.2059 -0.4150 0.2706 0.8188 1.4083

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.7777177 0.3821695 17.735 7.59e-13 \*\*\*

GrossUSD 0.0008011 0.0040185 0.199 0.844

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.127 on 18 degrees of freedom

Multiple R-squared: 0.002203, Adjusted R-squared: -0.05323

F-statistic: 0.03974 on 1 and 18 DF, p-value: 0.8442

**Regression for testing if all slopes are equal for the three genres:**

Analysis of Variance Table

Response: Rating

Df Sum Sq Mean Sq F value Pr(>F)

Genre 2 7.164 3.5820 3.7790 0.02910 \*

GrossUSD 1 2.340 2.3397 2.4683 0.12200

Genre:GrossUSD 2 6.303 3.1517 3.3250 0.04347 \*

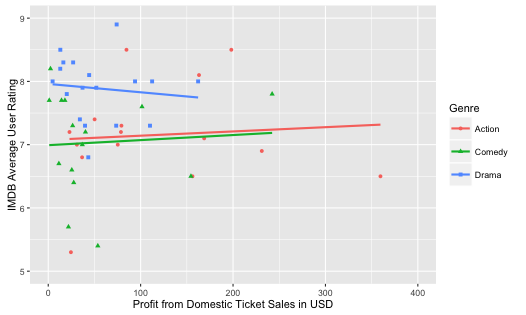
Residuals 54 51.186 0.9479

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Scatterplot of User Ratings vs. Profit, grouped by genre:**

Not all the slopes are zero, but profit is highly correlated to GrossUSD and will not be included.



[1] "Movie Genre Action"

Call:

lm(formula = Rating ~ profitUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-1.7444 -0.4741 0.1083 0.3777 1.6018

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.766202 0.257456 26.281 8.25e-16 \*\*\*

profitUSD 0.004111 0.001690 2.432 0.0257 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8877 on 18 degrees of freedom

Multiple R-squared: 0.2474, Adjusted R-squared: 0.2055

F-statistic: 5.916 on 1 and 18 DF, p-value: 0.02567

[1] "Movie Genre Drama"

Call:

lm(formula = Rating ~ profitUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-3.0992 -0.2838 0.1654 0.4773 1.3590

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.026869 0.326585 24.578 2.67e-15 \*\*\*

profitUSD -0.006561 0.004624 -1.419 0.173

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9504 on 18 degrees of freedom

Multiple R-squared: 0.1006, Adjusted R-squared: 0.05064

F-statistic: 2.013 on 1 and 18 DF, p-value: 0.173

[1] "Movie Genre Comedy"

Call:

lm(formula = Rating ~ profitUSD, data = sub)

Residuals:

Min 1Q Median 3Q Max

-2.9329 -0.6559 0.2007 0.8766 1.5568

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.633295 0.314457 21.094 4.2e-13 \*\*\*

profitUSD 0.004055 0.004147 0.978 0.343

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.136 on 16 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.05638, Adjusted R-squared: -0.002594

F-statistic: 0.956 on 1 and 16 DF, p-value: 0.3427

Due to similar results with GrossUSD, a correlation test was run with ProfitUSD and the two variables were 83% correlated. Therefore, ProfitUSD was not included at a covariate because it explains much of what GrossUSD explains.

Pearson's product-moment correlation

data: df$GrossUSD and df$profitUSD

t = 11.208, df = 56, p-value = 6.661e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.7302385 0.8972247

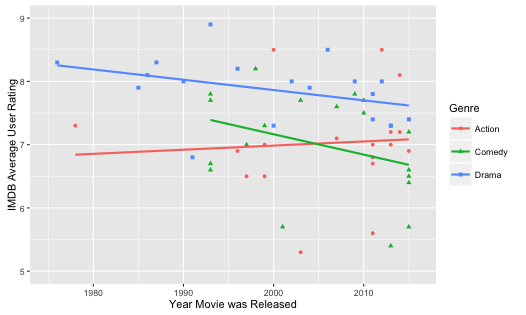
sample estimates:

cor

0.8316672

**Scatterplot of User Ratings vs. Year, grouped by genre:**

All the slopes are zero and Year will not be included as a covariate.



[1] "Movie Genre Action"

Call:

lm(formula = Rating ~ Year, data = sub)

Residuals:

Min 1Q Median 3Q Max

-1.8338 -0.4334 -0.1325 0.2156 2.6196

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11.547677 50.586546 -0.228 0.822

Year 0.009327 0.025213 0.370 0.716

Residual standard error: 1.019 on 18 degrees of freedom

Multiple R-squared: 0.007545, Adjusted R-squared: -0.04759

F-statistic: 0.1368 on 1 and 18 DF, p-value: 0.7158

[1] "Movie Genre Drama"

Call:

lm(formula = Rating ~ Year, data = sub)

Residuals:

Min 1Q Median 3Q Max

-3.0793 -0.1250 0.1141 0.3967 1.0076

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 77.26871 34.83524 2.218 0.0396 \*

Year -0.03478 0.01741 -1.998 0.0611 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9067 on 18 degrees of freedom

Multiple R-squared: 0.1815, Adjusted R-squared: 0.136

F-statistic: 3.991 on 1 and 18 DF, p-value: 0.06108

[1] "Movie Genre Comedy"

Call:

lm(formula = Rating ~ Year, data = sub)

Residuals:

Min 1Q Median 3Q Max

-3.3302 -0.3969 0.1793 0.7821 1.2293

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 47.42367 59.62976 0.795 0.437

Year -0.02025 0.02974 -0.681 0.505

Residual standard error: 1.114 on 18 degrees of freedom

Multiple R-squared: 0.02509, Adjusted R-squared: -0.02907

F-statistic: 0.4633 on 1 and 18 DF, p-value: 0.5047

**R code for the ANCOVA tests:**

##Setup working directory, import packages, and ingest data

setwd("~/Dropbox/Cuyler School/Applied Stats/STAT502/Project")

library(ggplot2)

library(multcomp)

df <- read.csv("./Project\_Data.csv")

#Rescaling GrossUSD and EstBudgetUSD to millions of dollars

df <- within(df, {

GrossUSD <- GrossUSD/1000000

EstBudgetUSD <- EstBudgetUSD/1000000

})

#Feature engineering for profitability

df$profitUSD <- df$GrossUSD-df$EstBudgetUSD

#ANCOVA Regression Tests for Differences in Genre

#Regression test for Rating vs. GrossUSD with plot

g <- ggplot(df, aes(y = Rating, x = GrossUSD, shape = Genre, color = Genre))

g <- g + geom\_point() + geom\_smooth(method="lm", fill=NA)

g <- g + xlab("Gross Domestic Ticket Sales in millions of USD")

g <- g + ylab("IMDB Average User Rating")

g <- g + ylim(5,9) + xlim(0, 550)

g

for(i in unique(df$Genre)){

print(paste("Movie Genre", i, sep = " "))

sub <- df[df$Genre == i,]

fit1 <- lm(Rating ~ GrossUSD, sub)

print(summary.lm(fit1))

}

#Due to the significant positive slope for Action movies and GrossUSD we test

#if all the slopes are equal with a significant interaction variable

df.anova1 <- aov(Rating ~ Genre \* GrossUSD, df)

anova(df.anova1)

#Regression test for Rating vs. profitUSD with plot

g <- ggplot(df, aes(y = Rating, x = profitUSD, shape = Genre, color = Genre))

g <- g + geom\_point() + geom\_smooth(method="lm", fill=NA)

g <- g + xlab("Profit from Domestic Ticket Sales in USD")

g <- g + ylab("IMDB Average User Rating")

g <- g + ylim(5,9) + xlim(0, 400)

g

for(i in unique(df$Genre)){

print(paste("Movie Genre", i, sep = " "))

sub <- df[df$Genre == i,]

fit1 <- lm(Rating ~ profitUSD, sub)

print(summary.lm(fit1))

}

#The results for profitUSD are similiar to GrossUSD and profitUSD so we tested

#for correlation and decided to drop profitUSD because it was 83% correlated to

#GrossUSD.

cor.test(df$GrossUSD, df$profitUSD)

#Regression test for Rating vs. Year with plot

g <- ggplot(df, aes(y = Rating, x = Year, shape = Genre, color = Genre))

g <- g + geom\_point() + geom\_smooth(method="lm", fill=NA)

g <- g + xlab("Year Movie was Released")

g <- g + ylab("IMDB Average User Rating")

g <- g + ylim(5,9) + xlim(1975, 2016)

g

for(i in unique(df$Genre)){

print(paste("Movie Genre", i, sep = " "))

sub <- df[df$Genre == i,]

fit1 <- lm(Rating ~ Year, sub)

print(summary.lm(fit1))

}

#There is one significant linear relationship between grossUSD and movie ratings

#and a significant interaction term so we proceed with an unequal slopes ANCOVA

#model

df.fit <- lm(Rating ~ Genre + Genre\* GrossUSD-1, df)

summary(df.fit)

#The follwing section presents the lsmeans difference in ratings for the three

#genres at three different levels of GrossUSD: 50, 150, and 250 million USD

levels <- data.frame(GrossUSD = seq(50, 250, 100))

fit<- lm(Rating ~ GrossUSD, df[df$Genre == "Action",])

action.ratings <- predict.lm(fit.action, newdata = levels)

fit <- lm(Rating ~ GrossUSD, df[df$Genre == "Drama",])

drama.ratings <- predict.lm(fit, newdata = levels)

fit <- lm(Rating ~ GrossUSD, df[df$Genre == "Comedy",])

comedy.ratings <- predict.lm(fit, newdata = levels)

print("The differences in the predicted average user rating between drama and action at the three levels:")

print(drama.ratings-action.ratings)

print("The differences in the predicted average user rating between drama and comedy at the three levels:")

print(drama.ratings-comedy.ratings)

print("The differences in the predicted average user rating between action and comedy at the three levels:")

print(action.ratings-comedy.ratings)

1. The formula that IMDB uses to select a “Random Title” is unknown. [↑](#footnote-ref-1)
2. IMDB Database Statistics. (21 April, 2016). Retrieved from http://www.imdb.com/stats [↑](#footnote-ref-2)