Rhea Qianqi Rao
Regression and Multivariate Data Analysis
Professor Simonoff
Nov 28th, 2015

Rotten Tomatoes Scores Of Movies – Do genres and movie time length matter?

A lot of previous researches were done examining the relationship between Box Office Revenue of a certain movie and its Rotten Tomatoes Scores or Revenue versus the duration of time since release date. I'm more interested in knowing the elements that affect the Rotten Tomatoes score for a certain movie, does the genre matter? What about the duration of the movie? Do audience has a general preference towards a certain genre? Are audience's preferences affected by how long the movies go on?

To answer these questions, I manually collected a sample size of 41 movie data from the Rotten Tomatoes site. I focus specifically on movies that were released after September 2015, and they are all under the category "Top Box Office" on the website. The ratings are shown in numerical numbers, so 0.50 translates to a 50% rating of the movie. To investigate, we try to fit an ANCOVA model.

I'll first examine the two independent variables more in depth:

- Runtime Of Movies (**TIME**): This refers to the duration of the specific movie in minutes. If we think about the relationship between TIME and the RT Score, we may expect a negative correlation since audience might get bored of the movie after a while.
- Genre Of Movies (**GEN**): This is our categorical variable, each of the 41 movies is classified as one of the 6 movie genres: Action, Drama, Comedy, Animation, Suspense, Science Fiction. It would be expected that different genres would have different preferences from the audience, but it is unclear which categories would have a more favorable impact on the movie rating.

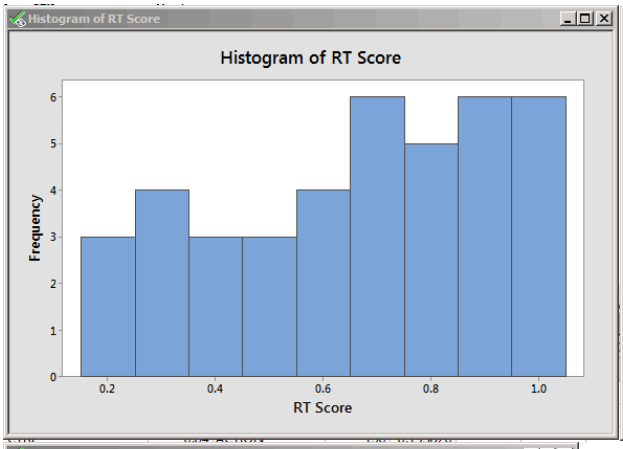
From Minitab, we obtained the descriptive statistics for RT Score, sorted by the GEN. Below we see that we have a quite unbalanced data, there is an obvious large number of Drama movies, while the others have similar counts.

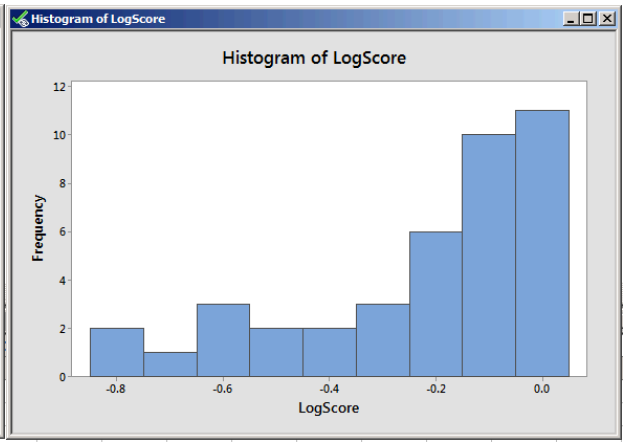
Variable	GEN	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
RT Score	ACTION	5	0	0.7480	0.0900	0.2013	0.4900	0.5650	0.7100	0.9500
	ANIMATION	5	0	0.7320	0.0896	0.2003	0.5200	0.5300	0.7600	0.9200
	COMEDY	4	0	0.438	0.113	0.225	0.210	0.228	0.440	0.645
	DRAMA	17	0	0.6871	0.0613	0.2529	0.2700	0.4100	0.7400	0.9350
	SCIENCE FICTION	4	0	0.535	0.191	0.382	0.160	0.185	0.525	0.895
	SUSPENSE	5	0	0.682	0.145	0.323	0.170	0.385	0.740	0.950

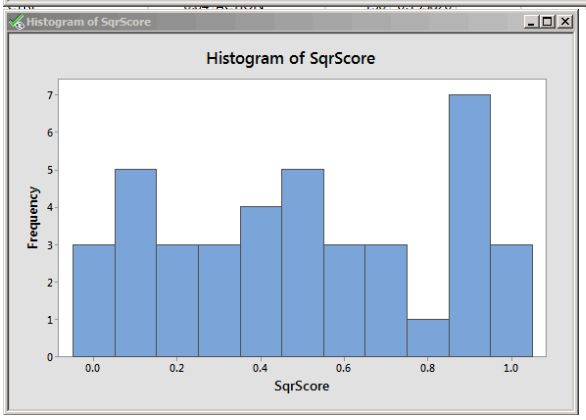
Variable	GEN	Maximum
RT Score	ACTION	0.9700
	ANIMATION	0.9800
	COMEDY	0.660
	DRAMA	0.9800
	SCIENCE FICTION	0.930
	SUSPENSE	0.980

On a first glance, different categories seem to have means close to each other, with Action having a more favorable rating and Comedy the lowest.

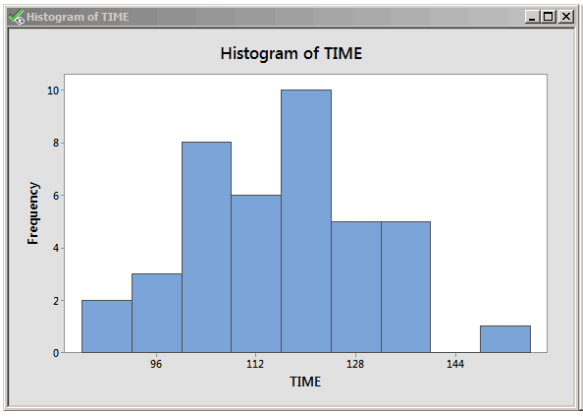
Since the histogram isn't particularly normal distributed, I tried taking the logs of and squaring the RT Score. However, the left tailed problem got even worse, so we stick with the original RT Score.

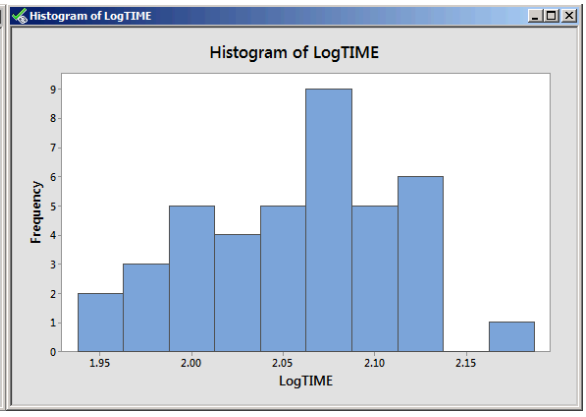




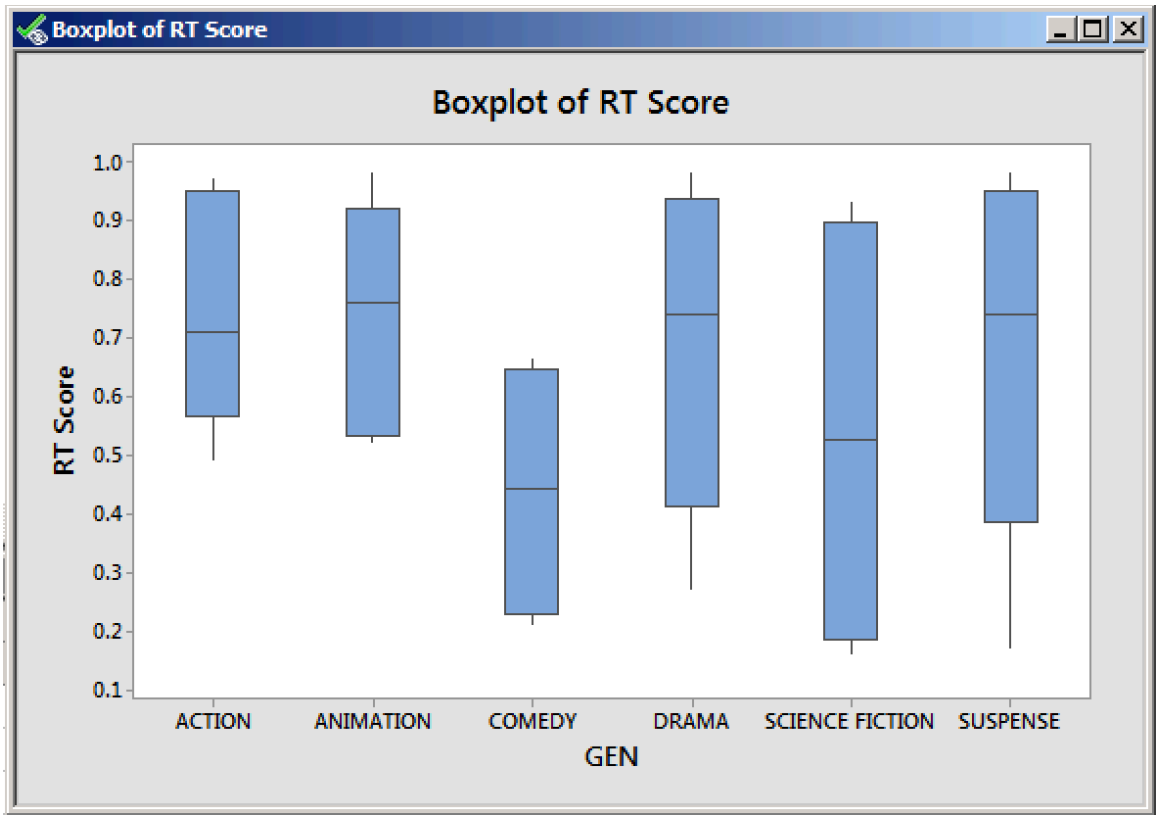


Examining TIME, the histogram is approximately normally distributed, it should be fine the way it is, but I took logTime just in case, and it turns out we should stick with the original TIME.



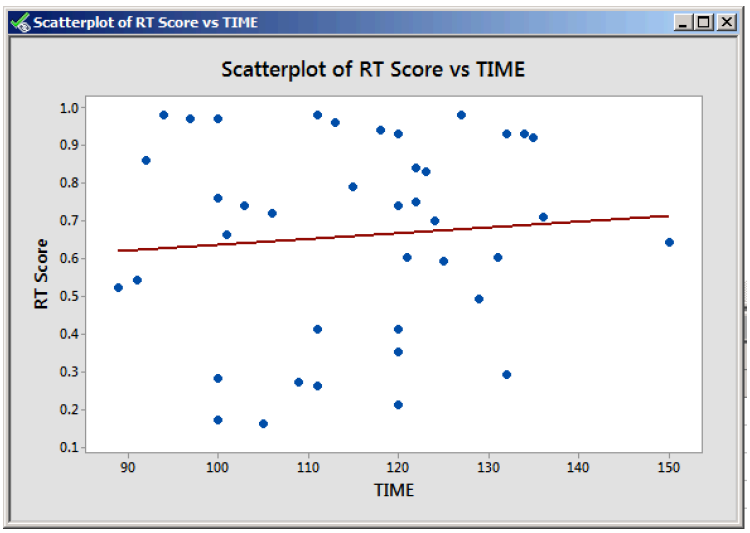


Let's now look at side-by-side boxplots of RT Score versus TIME. Below, we see not only are there differences in RT Scores between different genres, but also differences in variability, with Science Fiction have more variable ratings than other genres, and comedy having the least variability.

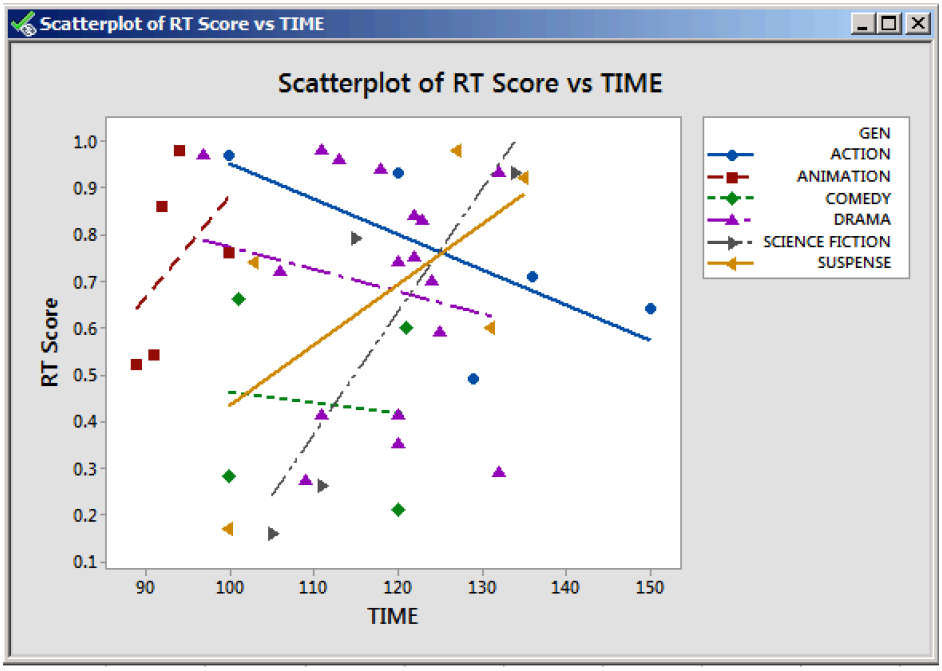


The relationship between the category of genres and the rating is ambiguous but worth exploring. It seems that the audience is not particularly favorable towards any type of genre of movies, but they generally give a lower rating to comedies and Science Fiction movies. One guess could be that Comedies are usually not super impressive work like the Inception, they give the audience a good laugh, but they tend not to be particularly memorable. We also see a bigger variability in Science Fiction, which could translate to meaning that Science Fiction is usually a hit or miss, a very risky genre to be filming with. It's hard to be certain about why we observe these data, but it's interesting to postulate.

Next we take a look at the scatterplot of our numerical predictor (TIME) against RT Score. The plot shows arguably a very weak positive relationship between RT Score and TIME, if at all. It really just seems like there isn't quite a correlation between the response variable and TIME.



What happens when we separate the data points by their GEN category? We turn to the scatterplot below, where we find that the relationship between time and RT Score is different for every genre. The diverse regression lines demonstrate that given the genre, RT Score is correlated to the time length of the movie in different ways, and Animation as a group has a much shorter time than average. This is an interesting thing to note about the data, but first we'll proceed with ANCOVA testing.



Generating the ANCOVA output in Minitab, we get:

General Linear Model: RT Score versus TIME, GEN

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
GEN	Fixed	6	ACTION, ANIMATION, COMEDY, DRAMA, SCIENCE FICTION, SUSPENSE

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
TIME	1	0.02115	0.02115	0.30	0.588
GEN	5	0.34156	0.06831	0.97	0.452
Error	33	2.33222	0.07067		
Lack-of-Fit	28	1.87272	0.06688	0.73	0.736
Pure Error	5	0.45950	0.09190		
Total	39	2.69339			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.265845	13.41%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.396	0.442	0.90	0.377	
TIME	0.00211	0.00386	0.55	0.588	1.72
GEN					
ACTION	0.084	0.119	0.70	0.488	2.01
ANIMATION	0.139	0.135	1.03	0.310	2.57
COMEDY	-0.192	0.119	-1.61	0.117	1.81
DRAMA	0.0418	0.0726	0.58	0.568	1.37
SCIENCE FICTION	-0.107	0.119	-0.90	0.376	1.79

Regression Equation

GEN

ACTION RT Score = 0.480 +†0.00211†TIME

ANIMATION RT Score = 0.535 +†0.00211†TIME

COMEDY RT Score = 0.204 +†0.00211†TIME

DRAMA RT Score = 0.438 +†0.00211†TIME

SCIENCE FICTION RT Score = 0.290 +†0.00211†TIME

SUSPENSE RT Score = 0.430 +†0.00211†TIME

Fits and Diagnostics for Unusual Observations

```
Obs  RT Score    Fit   Resid  Std Resid
 35     0.170  0.641  -0.471      -2.09  R

R  Large residual

Means

                   Fitted
Term                 Mean  SE Mean
GEN
  ACTION            0.723    0.127
  ANIMATION         0.778    0.146
  COMEDY            0.447    0.134
  DRAMA            0.6813   0.0653
  SCIENCE FICTION   0.533    0.133
  SUSPENSE          0.674    0.120


Covariate  Data Mean  StDev
TIME           115.2   14.5
```

The model has an R^2 of 13.41%, not exactly great news. We see that neither the duration of the movie nor the genre of the movie is statistically significant predictors of RT Score. GEN has a F-statistic of 0.97 and a P-value of 0.452, so given TIME, there really isn't evidence to reject the statement that there isn't significant difference between the GEN groups. Same goes for the TIME variable, which has an even bigger P-value.

Nevertheless, we still want to know what the data means – even if they are not statistically significant, so we have an overall level of RT Score of 0.396. Relative to overall level, Comedy has the biggest negative effect compared to the overall level and Animation the biggest positive one. PI is 0.52, which is half of 100%, considering that we have decided to not take log of RT Score, it really isn't a helpful figure.

When we look at the entries under **Term**. The data indicate that for a given genre, there is a positive correlation between RT Score and TIME. This is in line with the scatterplot we did earlier of RT Rating Versus TIME, but not in line with our initial hypothesis. The data is maybe telling us that given a certain genre, the longer the movie is, the more the audience is expected to enjoy it.

What do the fitted means output tell us? These can help us directly interpret what's going on in the model. For example, looking at Drama, the estimated expected RT Score is 68.1% when all the numerical predictor (TIME) is equal to its mean. It also allows you to see the differences between groups. Keeping TIME fixed, Animation has a higher expected rating than those from the other genres. The estimated difference in the mean between Animation and Comedy is (0.778 -0.447 = 0.331) given a fixed number of TIME.

The fitted means deserve further comment. The estimated means for RT Score for the genre of the movie:

Descriptive Statistics: RT Score

```
Variable  GEN              N    Mean  SE Mean  TrMean   StDev  Median
RT Score  ACTION           5  0.7480   0.0900       *  0.2013  0.7100
          ANIMATION        5  0.7320   0.0896       *  0.2003  0.7600
```

```
COMEDY            4   0.438    0.113       *   0.225   0.440
DRAMA            17  0.6871   0.0613  0.6953  0.2529  0.7400
SCIENCE FICTION   4   0.535    0.191       *   0.382   0.525
SUSPENSE          5   0.682    0.145       *   0.323   0.740
```

The fitted mean is generally slightly smaller than the unadjusted mean, except for Comedy. This is probably because Comedy in general as a group has a lower than average TIME. However the overall pattern is not that different.

We have yet to consider a potential interaction effect here. We add the interaction of GEN with TIME and obtain the following ANCOVA output:

General Linear Model: RT Score versus TIME, GEN

```
Method

Factor coding  (-1, 0, +1)


Factor Information

Factor  Type   Levels  Values
GEN     Fixed       6  ACTION, ANIMATION, COMEDY, DRAMA, SCIENCE FICTION,
SUSPENSE


Analysis of Variance

Source          DF  Adj SS   Adj MS  F-Value  P-Value
  TIME           1  0.1019  0.10187     1.68    0.206
  GEN            5  0.6552  0.13103     2.16    0.088
  TIME*GEN       5  0.6318  0.12636     2.08    0.098
Error           28  1.7004  0.06073
  Lack-of-Fit   23  1.2409  0.05395     0.59    0.826
  Pure Error     5  0.4595  0.09190
Total           39  2.6934


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.246433  36.87%     12.06%       0.00%


Coefficients

Term                  Coef  SE Coef  T-Value  P-Value   VIF
Constant             0.396    0.442     0.90    0.377
TIME               0.00211  0.00386     0.55    0.588  1.72
GEN
  ACTION             0.084    0.119     0.70    0.488  2.01
  ANIMATION          0.139    0.135     1.03    0.310  2.57
  COMEDY            -0.192    0.119    -1.61    0.117  1.81
  DRAMA             0.0418   0.0726     0.58    0.568  1.37
  SCIENCE FICTION   -0.107    0.119    -0.90    0.376  1.79


Regression Equation

GEN
```

```
ACTION           RT Score = 0.480 +†0.00211†TIME

ANIMATION        RT Score = 0.535 +†0.00211†TIME

COMEDY           RT Score = 0.204 +†0.00211†TIME

DRAMA            RT Score = 0.438 +†0.00211†TIME

SCIENCE FICTION  RT Score = 0.290 +†0.00211†TIME

SUSPENSE         RT Score = 0.430 +†0.00211†TIME


Fits and Diagnostics for Unusual Observations

Obs  RT Score    Fit   Resid  Std Resid
 35     0.170  0.641  -0.471      -2.09  R
```

Assuming a confidence level of 95%, we can see that while adding the interaction of the two variables might be helpful, the p-value is still statistically insignificant, the P-value is indicating there is a different line for different genres, as we anticipated earlier in our analysis. While our R^2 improved to 36.8%, the interaction variable has a statistically insignificant t-test. So I think it is a better idea to exclude the interaction component.

Since we opted for an ANCOVA model that did not include an interaction, we can perform a Tukey comparisons test to see which categories are significantly different from each other:

Comparisons for RT Score

Tukey Pairwise Comparisons: Response = RT Score, Term = GEN

```
Grouping Information Using the Tukey Method and 95% Confidence

GEN              N      Mean  Grouping
ACTION           5  0.748000  A
ANIMATION        5  0.732000  A
DRAMA        |  17  0.687059  A
SUSPENSE         5  0.682000  A
SCIENCE FICTION  4  0.535000  A
COMEDY           4  0.437500  A

Means that do not share a letter are significantly different.
```

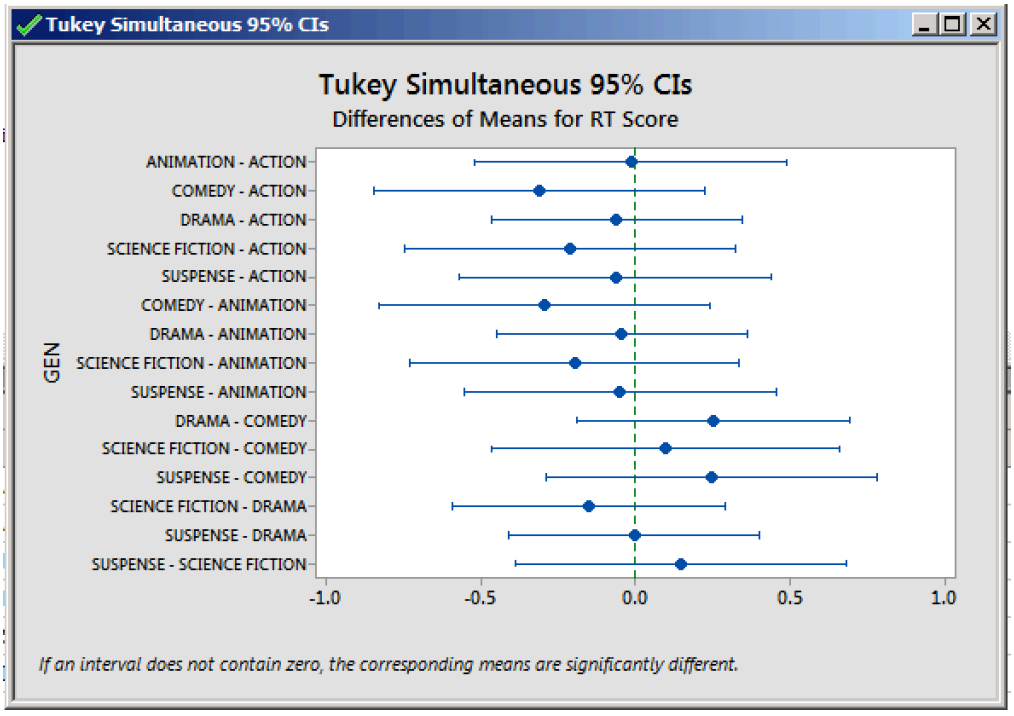
For my data, none of the groups were that different from each other, they all kinda overlap with each other. Adjusted P-value is basically telling us how strong the relationships are. None of the P-value here is really statistically significant.

Tukey Simultaneous Tests for Differences of Means

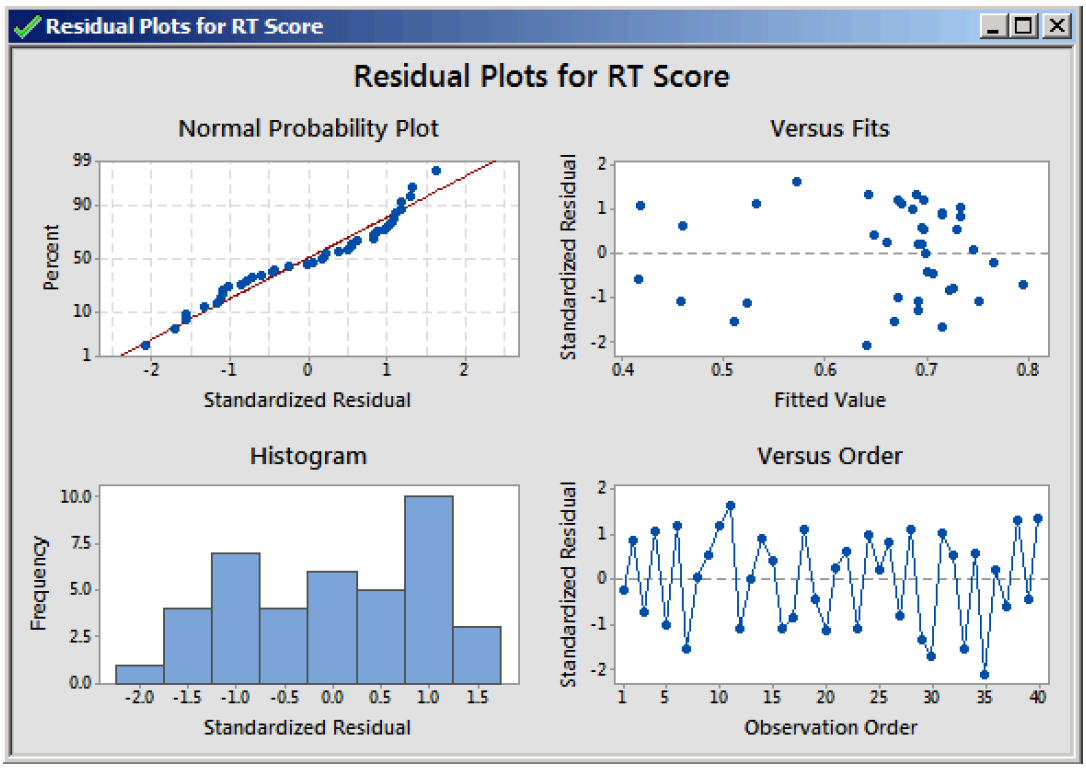
Difference of GEN Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
ANIMATION - ACTION	-0.016	0.166	(-0.518, 0.486)	-0.10	1.000
COMEDY - ACTION	-0.311	0.176	(-0.843, 0.222)	-1.76	0.504
DRAMA - ACTION	-0.061	0.134	(-0.465, 0.343)	-0.46	0.997
SCIENCE FICTION - ACTION	-0.213	0.176	(-0.746, 0.320)	-1.21	0.830
SUSPENSE - ACTION	-0.066	0.166	(-0.568, 0.436)	-0.40	0.999
COMEDY - ANIMATION	-0.295	0.176	(-0.827, 0.238)	-1.67	0.561
DRAMA - ANIMATION	-0.045	0.134	(-0.449, 0.359)	-0.34	0.999
SCIENCE FICTION - ANIMATION	-0.197	0.176	(-0.730, 0.336)	-1.12	0.871
SUSPENSE - ANIMATION	-0.050	0.166	(-0.552, 0.452)	-0.30	1.000
DRAMA - COMEDY	0.250	0.146	(-0.192, 0.691)	1.71	0.537
SCIENCE FICTION - COMEDY	0.098	0.186	(-0.464, 0.659)	0.52	0.995
SUSPENSE - COMEDY	0.245	0.176	(-0.288, 0.777)	1.39	0.735
SCIENCE FICTION - DRAMA	-0.152	0.146	(-0.594, 0.289)	-1.04	0.901
SUSPENSE - DRAMA	-0.005	0.134	(-0.409, 0.399)	-0.04	1.000
SUSPENSE - SCIENCE FICTION	0.147	0.176	(-0.386, 0.680)	0.83	0.959

Individual confidence level = 99.52%

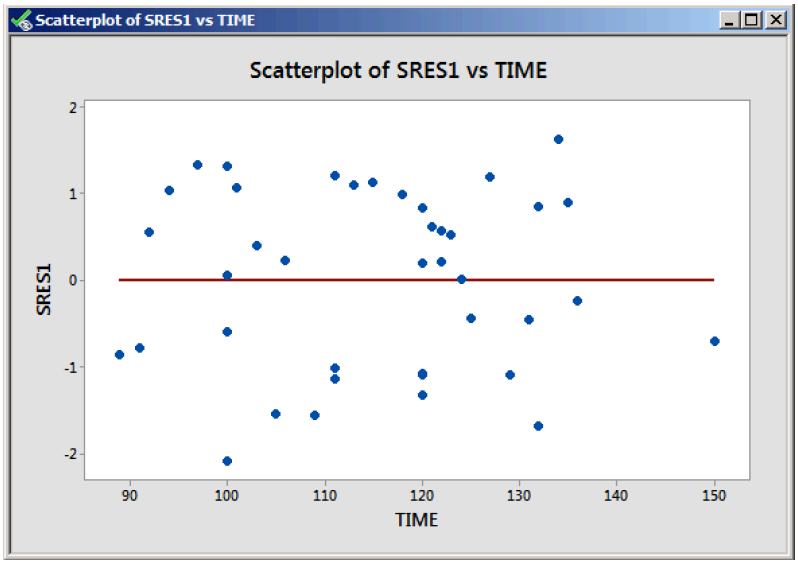
Tukey Simultaneous 95% CIs



As the descriptive statistic and boxplots suggested, the 6 genres of movies are not that distinct from each other with respect to RT Score. Now let's check to make sure that the models' assumptions are not violated. Given that our data set is not time-series based, we do not need to worry about autocorrelation. Let's take a look at the residuals plots.

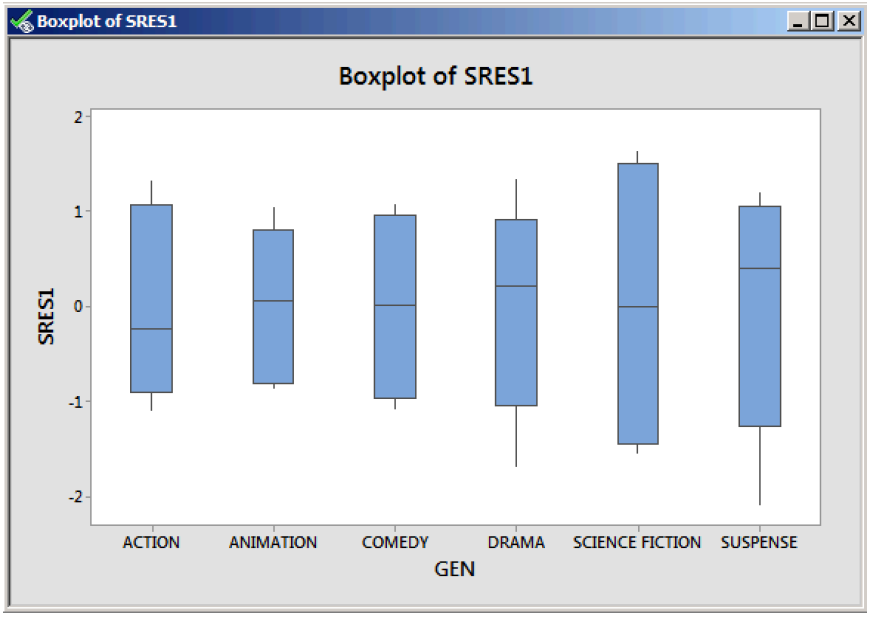


From the Normal Probability Plot and the Histogram, the residuals look approximately normally distributed, although not obviously so. Turning to the Residuals versus Fits Plot, although the order graph seems evenly distributed around 0, we see an obvious clustering around 0.7, which could be an evidence for non-constant variance. So we proceed to further examine residuals with respect to TIME and GEN by generating a scatterplot of residuals versus TIME.

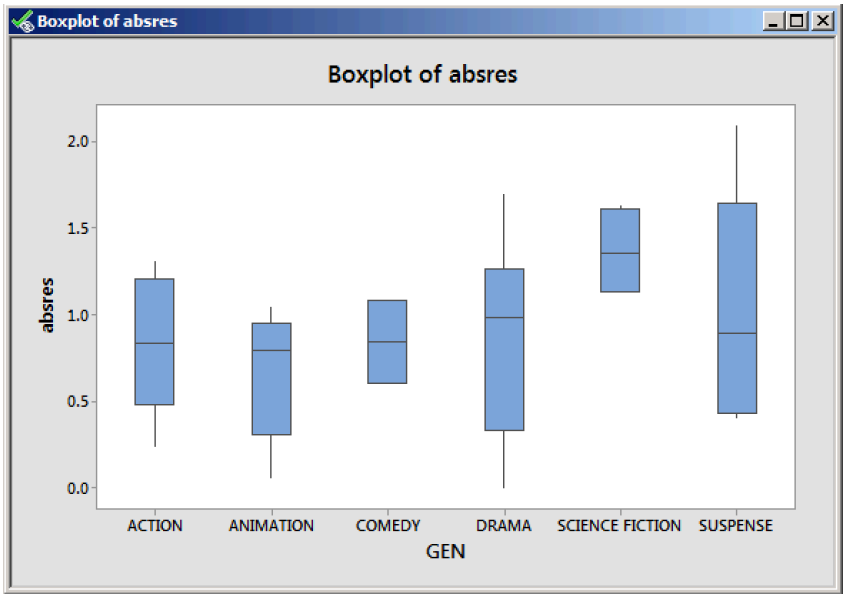


Looking at Residual Versus TIME, there does not seem to be evidence of non-constant variance with respect to TIME. Next, we can look for evidence of non-constant variance amongst the GEN by generating side-by-side boxplots of the residuals for each genre. Below,

we do see signs of non-constant variance – although not that obvious. Science Fiction has the greatest variability in RT Rating, followed by Suspense and Action, which are pretty close to each other in terms of variability.



The next logical step would be Levene's Test (and Weighted Least Squares if non-constant variance is detected). A Levene's test applied to the data below confirms the constant variance:



General Linear Model: absSRES1 versus GEN

```
Method

Factor coding   (-1, 0, +1)
```

```
Factor Information

Factor  Type   Levels  Values
GEN     Fixed       6  ACTION, ANIMATION, COMEDY, DRAMA, SCIENCE FICTION,
SUSPENSE


Analysis of Variance

Source  DF  Adj SS  Adj MS  F-Value  P-Value
  GEN    5   1.294  0.2588     1.12    0.366
Error   34   7.825  0.2301
Total   39   9.119


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.479739  14.19%      1.57%       0.00%
```

With P-value of 0.366, the data seems to tell us that there isn't a violation of assumptions, and that there is constant variance. After double-checking the P-Value is correct, I decided that I don't need to run WLS model.

Since we opted for an ANCOVA model that did not include an interaction, we can perform a Tukey comparisons test to see which categories are significantly different from each other:

Comparisons for RT Score

Tukey Pairwise Comparisons: Response = RT Score, Term = GEN

Grouping Information Using the Tukey Method and 95% Confidence

GEN	N	Mean	Grouping
ACTION	5	0.748000	A
ANIMATION	5	0.732000	A
DRAMA	17	0.687059	A
SUSPENSE	5	0.682000	A
SCIENCE FICTION	4	0.535000	A
COMEDY	4	0.437500	A

Means that do not share a letter are significantly different.

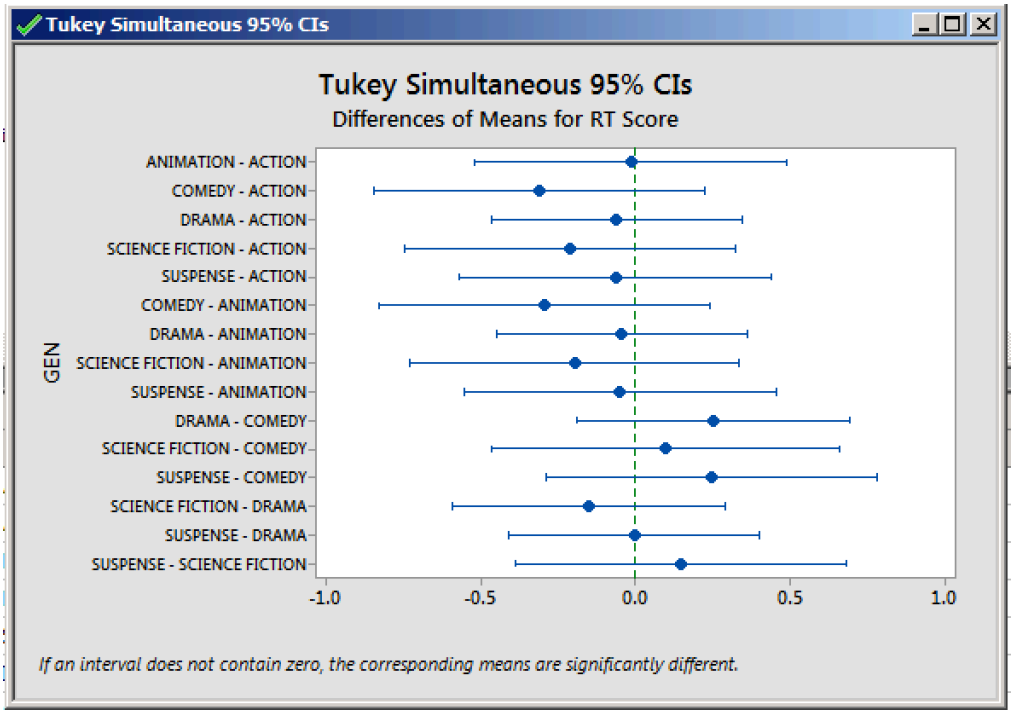
For my data, none of the groups were that different from each other, they all kinda overlap with each other. Adjusted P-value is basically telling us how strong the relationships are. None of the P-value here is really statistically significant.

Tukey Simultaneous Tests for Differences of Means

Difference of GEN Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
ANIMATION - ACTION	-0.016	0.166	(-0.518, 0.486)	-0.10	1.000
COMEDY - ACTION	-0.311	0.176	(-0.843, 0.222)	-1.76	0.504
DRAMA - ACTION	-0.061	0.134	(-0.465, 0.343)	-0.46	0.997
SCIENCE FICTION - ACTION	-0.213	0.176	(-0.746, 0.320)	-1.21	0.830
SUSPENSE - ACTION	-0.066	0.166	(-0.568, 0.436)	-0.40	0.999
COMEDY - ANIMATION	-0.295	0.176	(-0.827, 0.238)	-1.67	0.561
DRAMA - ANIMATION	-0.045	0.134	(-0.449, 0.359)	-0.34	0.999
SCIENCE FICTION - ANIMATION	-0.197	0.176	(-0.730, 0.336)	-1.12	0.871
SUSPENSE - ANIMATION	-0.050	0.166	(-0.552, 0.452)	-0.30	1.000
DRAMA - COMEDY	0.250	0.146	(-0.192, 0.691)	1.71	0.537
SCIENCE FICTION - COMEDY	0.098	0.186	(-0.464, 0.659)	0.52	0.995
SUSPENSE - COMEDY	0.245	0.176	(-0.288, 0.777)	1.39	0.735
SCIENCE FICTION - DRAMA	-0.152	0.146	(-0.594, 0.289)	-1.04	0.901
SUSPENSE - DRAMA	-0.005	0.134	(-0.409, 0.399)	-0.04	1.000
SUSPENSE - SCIENCE FICTION	0.147	0.176	(-0.386, 0.680)	0.83	0.959

Individual confidence level = 99.52%

Tukey Simultaneous 95% CIs



As the descriptive statistic and boxplots suggested, the 6 genres of movies are not that distinct from each other with respect to RT Score. Now let's check for abnormal points.

Row	Movie	SRES1	HI1	COOK1
1	MOCKING JAY	-0.24227	0.217052	0.002325
2	CREED	0.84586	0.100432	0.011411
3	SPECTRE	-0.70953	0.311362	0.032518

4	THE NIGHT BEFORE	1.0671	0.268999	0.059861
5	SECRET IN THEIR EYES	-1.023	0.068966	0.011074
6	BROOKLYN	1.19911	0.068966	0.015216
7	VICTOR FRANKENSTEIN	-1.55793	0.075653	0.028379
8	THE GOOD DINOSAUR	0.05777	0.209734	0.000127
9	THE PEANUTS MOVIE	0.54907	0.200303	0.010787
10	SPOTLIGHT	1.19364	0.212808	0.055025
11	THE MARTIAN	1.6266	0.316325	0.174883
12	LOVE THE COOPERS	-1.0891	0.268999	0.062355
13	TRUMBO	0.00061	0.066551	0
14	BRIDGE OF SPIES	0.89047	0.252553	0.038275
15	GOOSEBUMPS	0.4018	0.255247	0.007904
16	THE 33	-1.09162	0.059716	0.010811
17	HOTEL TRANSYLVANIA 2	-0.85631	0.203713	0.026798
18	ROOM	1.10173	0.063963	0.011849
19	LEGEND	-0.45804	0.229312	0.008918
20	PAN	-1.15081	0.255802	0.065032
21	SUFFRAGETTE	0.2291	0.088841	0.000731
22	THE INTERN	0.61925	0.273209	0.020593
23	MAZE RUNNER	-1.10337	0.200842	0.043709
24	CAROL	0.98026	0.058824	0.00858
25	THE DANISH GIRL	0.18852	0.059716	0.000322
26	SICARIO	0.83292	0.210315	0.026395
27	MINIONS	-0.78845	0.201019	0.022344
28	ANTMAN	1.1193	0.250329	0.059763
29	WAR ROOM	-1.32437	0.059716	0.015913
30	BY THE SEA	-1.69239	0.100432	0.045681
31	INSIDE OUT	1.03598	0.200135	0.038362
32	WOODLAWN	0.51432	0.064211	0.002593
33	THE LAST WITCH HUNTER	-1.55356	0.276643	0.131864
34	STEVE JOBS	0.56084	0.062292	0.002985
35	THE PERFECT GUY	-2.08668	0.277604	0.239035
36	BLACK MASS	0.21123	0.062292	0.000423
37	BURNT	-0.59719	0.273209	0.019152
38	THEEB	1.30507	0.353464	0.133021
39	TRUTH	-0.43652	0.069313	0.002027
40	MUSTANG	1.33558	0.151141	0.045372

Looking at the hat value and cook's number, we'll see that none is particularly statistically significant. There also isn't any leverage points that have hii > 0.427. So there really isn't an outlier.

Conclusion

After a thorough process of examining the regression assumptions and diagnostics, and testing different ANCOVA models, we learn that there aren't any solid correlation between the Rotten Tomatoes Score of a movie, and the genre of that specific movie, nor is there a solid correlation between the time length of a movie and its rating. Although we interestingly noted that certain genre of the movie have shorter movie time length as a group in our sample, but it is also important to note that we have 41 samples in total, and that our sample size might not be big enough for us to make any claims.

When we tested a model with different slopes, we found it was not significantly better than a model with the original slope. Ultimately, we elected to choose an ANCOVA model.