“SHOULD I WATCH THIS HORROR MOVIE?”—HOW TO PREDICT IMDB SCORES OF HORROR MOVIES IF YOU ARE HESITATING, USING SAS

By Jing Kunzler for OPR 9750 Final Project

EXECUTIVE SUMMARY

This project discusses the popular opinion that most horror movies are disappointing, factors we assume contribute to that opinion, and build the best model to predict the rating of horror movies, for those who have a list of “maybe, maybe not” in mind, who wonder what ways can we quantify and calculate a relatively fair score, and for those who have a general interest in understanding horror movies from a quantitative perspective.

Since the first horror movie created in late 1890s, the depiction of human’s fantasy of fear, has steadily become a solid mainstream genre. Being a fan of thriller/supernatural themes, I like horror movies that mean more than gore and scare, but reflect the real fear as the society changes, like Frankenstein or vampire themes that reflected the confusion of social changes, science, and religion of the industrial era. But more horror films of the modern era have been disappointing, and more people wonder why they still keep coming out.

The dataset used to study was found on Data.world, put together by Chuan Sun (@sundeepblue on Github) who scrapped metadata using a combination of www.the-numbers.com, IMDB.com, and a Python library called "scrapy". This data set includes six regressors and one response variable. The six regressors are number of critics reviewed, duration (in minutes), domestic gross, number of imdb users voted, number of users reviewed, and budget of each film. And the response variable, the goal of our prediction, is imdb scores.

The analysis is divided by a four-part analysis consists of testing initial model and regression assumptions, model selection, individual response prediction intervals, and conclusion.

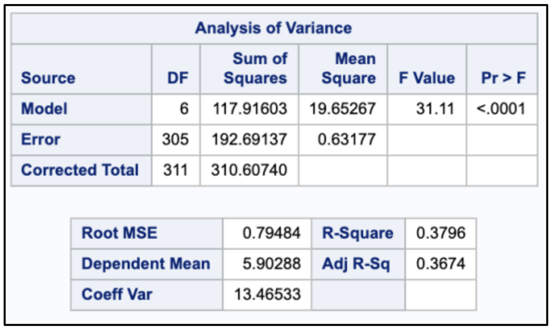
INITIAL MODEL & REGRESSION ASSUMPTIONS

First, we have to prove that there is linear relationship between each predictor variables—number of critics reviewed, duration (in minutes), domestic gross, number of imdb users voted, number of users reviewed, and budget of each film, and the response variable—imdb scores. Since domestic gross was transformed into millions, budget was transformed into millions as well.

We will run initial modeling to check the 4 assumptions of linear regression. These assumptions are:

1. Multicollinearity
2. Normality
3. Linearity
4. Homoscedasticity (equal variance across observations)

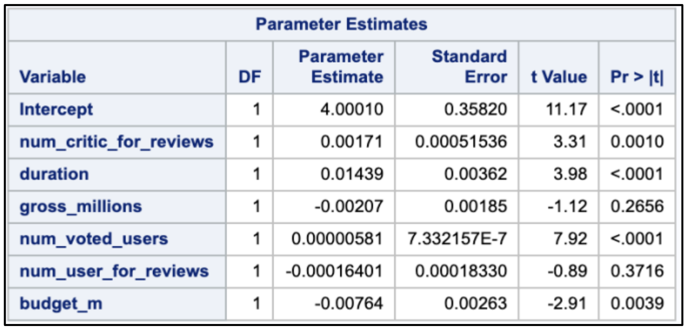
An initial model that contains all the predictors against the response variable shows the below results:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	117.91603	19.65267	31.11	<.0001
Error	305	192.69137	0.63177		
Corrected Total	311	310.60740			

Root MSE	0.79484	R-Square	0.3796
Dependent Mean	5.90288	Adj R-Sq	0.3674
Coeff Var	13.46533		



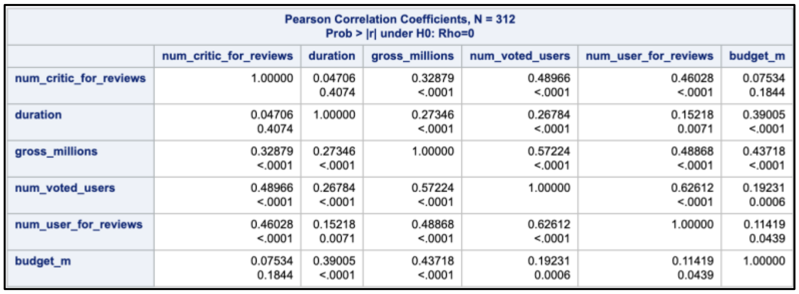
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.00010	0.35820	11.17	<.0001
num_critic_for_reviews	1	0.00171	0.00051536	3.31	0.0010
duration	1	0.01439	0.00362	3.98	<.0001
gross_millions	1	-0.00207	0.00185	-1.12	0.2656
num_voted_users	1	0.00000581	7.332157E-7	7.92	<.0001
num_user_for_reviews	1	-0.00016401	0.00018330	-0.89	0.3716
budget_m	1	-0.00764	0.00263	-2.91	0.0039

While the model is significant, suggested by the p-value at 95% confidence of the overall model—meaning that at least one of these predictors is useful to predict imdb scores, only the duration and number of of imdb users reviewed are useful to predict imdb scores. We question that other predictors such as domestic gross or budget are also useful, considering how big commercial productions usually attract more attention, which might lead to more critics or users reviewing them.

I. Multicollinearity

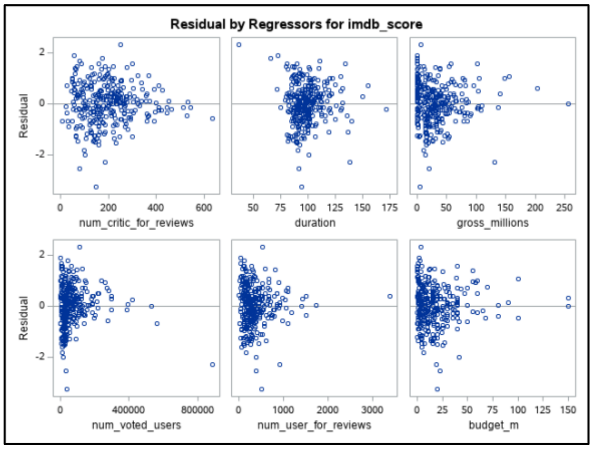
Multicollinearity means one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. We don't want to include variables that are highly correlated with each other in the same model because our data will be biased, and model, inaccurate. We checked multicollinearity by running a CORR procedure to check significant correlations (p-value < 0.001).

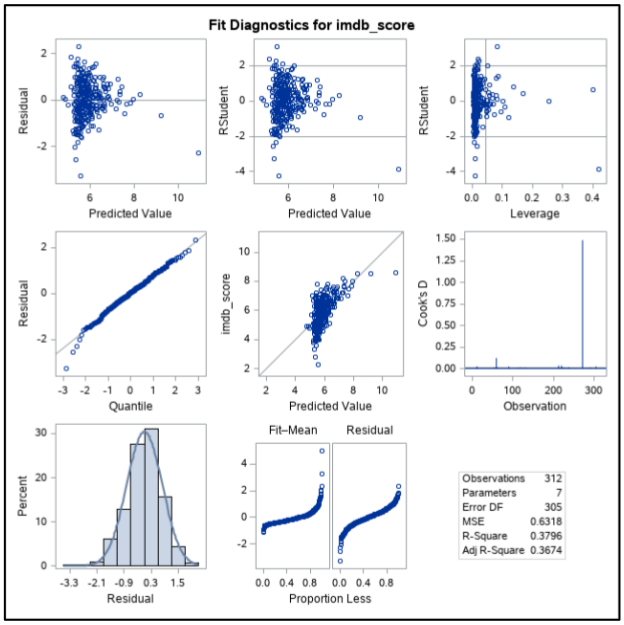


Pearson Correlation Coefficients, N = 312
Prob > |r| under H0: Rho=0

	num_critic_for_reviews	duration	gross_millions	num_voted_users	num_user_for_reviews	budget_m
num_critic_for_reviews	1.00000	0.04706 0.4074	0.32879 <.0001	0.48966 <.0001	0.46028 <.0001	0.07534 0.1844
duration	0.04706 0.4074	1.00000	0.27346 <.0001	0.26784 <.0001	0.15218 0.0071	0.39005 <.0001
gross_millions	0.32879 <.0001	0.27346 <.0001	1.00000	0.57224 <.0001	0.48868 <.0001	0.43718 <.0001
num_voted_users	0.48966 <.0001	0.26784 <.0001	0.57224 <.0001	1.00000	0.62612 <.0001	0.19231 0.0006
num_user_for_reviews	0.46028 <.0001	0.15218 0.0071	0.48868 <.0001	0.62612 <.0001	1.00000	0.11419 0.0439
budget_m	0.07534 0.1844	0.39005 <.0001	0.43718 <.0001	0.19231 0.0006	0.11419 0.0439	1.00000

II. Normality & Linearity & Homoscedasticity



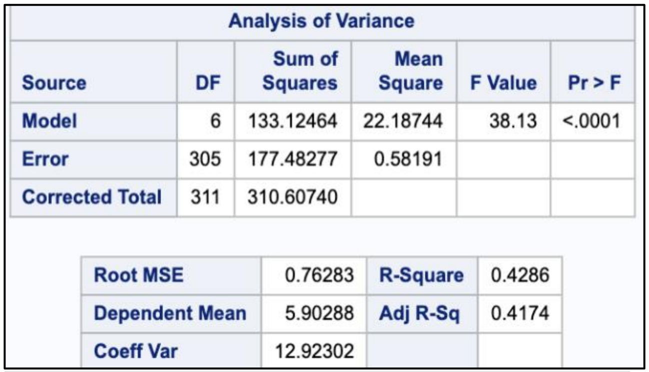


Observations	312
Parameters	7
Error DF	305
MSE	0.6318
R-Square	0.3796
Adj R-Square	0.3674

In the second picture, we can see that the upper left corner residuals plot doesn't show a "nonrandom" pattern, which means that the distribution of residuals isn't normal. In the same second picture above, the observed and predicted imdb scores don't line up roughly on the linear line and appears to be curved to the southeast direction.

In the picture on the left above, we can see that the residuals of each predictor against the response variable doesn't appear "nonrandom". They all have a cone like pattern, with the exception of first plot slightly more random than the others.

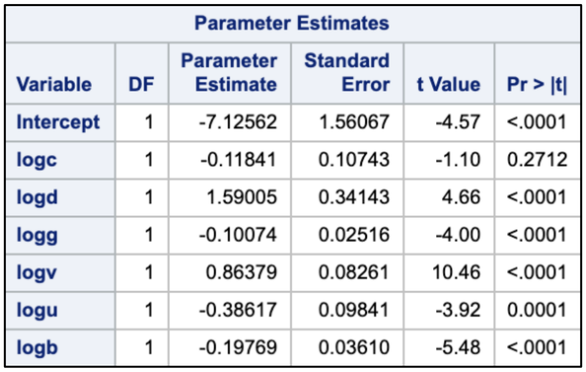
We used log transformation to resolve the assumptions issues.



Analysis of Variance

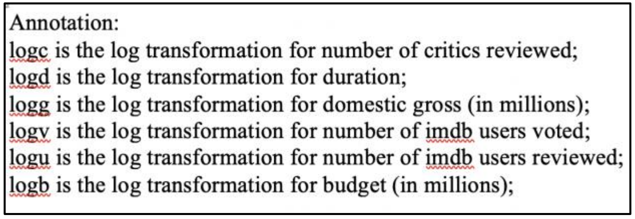
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	133.12464	22.18744	38.13	<.0001
Error	305	177.48277	0.58191		
Corrected Total	311	310.60740			

Root MSE	0.76283	R-Square	0.4286
Dependent Mean	5.90288	Adj R-Sq	0.4174
Coeff Var	12.92302		



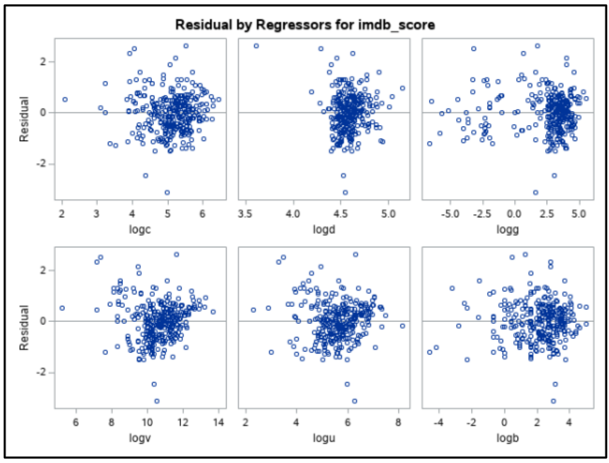
Parameter Estimates

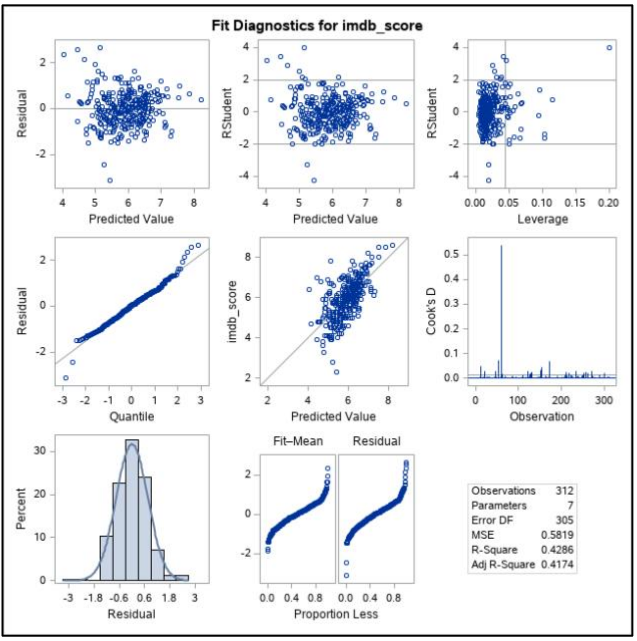
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.12562	1.56067	-4.57	<.0001
logc	1	-0.11841	0.10743	-1.10	0.2712
logd	1	1.59005	0.34143	4.66	<.0001
logg	1	-0.10074	0.02516	-4.00	<.0001
logv	1	0.86379	0.08261	10.46	<.0001
logu	1	-0.38617	0.09841	-3.92	0.0001
logb	1	-0.19769	0.03610	-5.48	<.0001



Annotation:
logc is the log transformation for number of critics reviewed;
logd is the log transformation for duration;
logg is the log transformation for domestic gross (in millions);
logv is the log transformation for number of imdb users voted;
logu is the log transformation for number of imdb users reviewed;
logb is the log transformation for budget (in millions);

The new significant predictors are the log of, duration, domestic gross, number of imdb users voted, and budget, which definitely makes more sense given the previous thinking that domestic gross and budget should impact imdb scores.





The assumptions are met this time. 1) The first residual plot on the second picture shows a "random pattern"—this meets the normality of the residuals assumption. 2) The data points gather tightly along the y = x linearity line, on the middle plot of the second row, on the second picture. This meets the linearity assumption. 3) The residuals plots of the predictors against imdb scores on the first picture all show no pattern. This means equal variance, or Homoscedasticity is met.

Since this is an initial modeling to understand and prepare appropriate data that meets regression assumptions, we still need to go through model selection that picks the most suitable model.

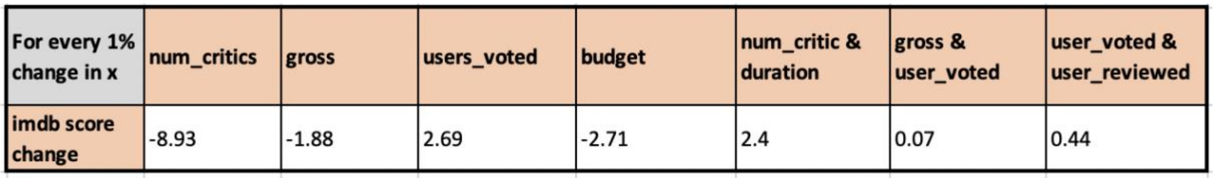
MODEL SELECTION

There are several methods of model selection. There are three types based on how to introduce or eliminate variables, including forward, backward, and stepwise. Forward selection begins with an empty model, adds a variable until the best improvement is met. Backward selection starts with all the variables, and each step eliminates the most useless variable. Stepwise simultaneously add or eliminate variables as needed.

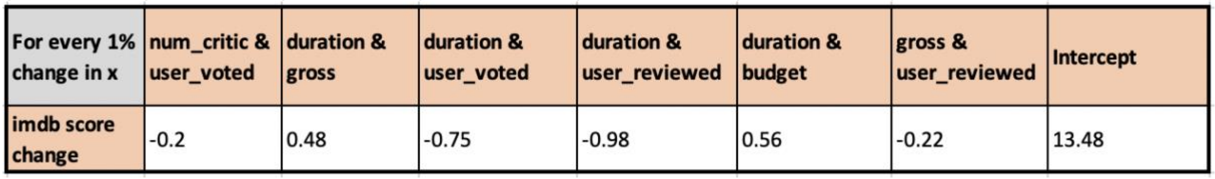
Besides the selection methods, there are many statistics that can be used as selection criterions for comparison. P-value, adjusted R-square, Mallow's cp, AIC, BIC…etc. Many people prefer some over the others for various reasons, as they have unique advantages and flaws. Today we will use Mallow's cp as criterion and go through all three selection methods to that has the smallest cp value. We also included interaction terms between each pair of the predictors to inspect if interaction between variables improve or worsen our model.

The smallest cp, i.e. the tightest fit model was generated by backward selection at 11.99961. Next best is forward, and stepwise selection produced the worst result.

Now we can write the equation. The equation is long and overwhelming to write out, because of all the interaction terms and the log transformation, so please see the tables below that explains the coefficients for the log-transformed variables.



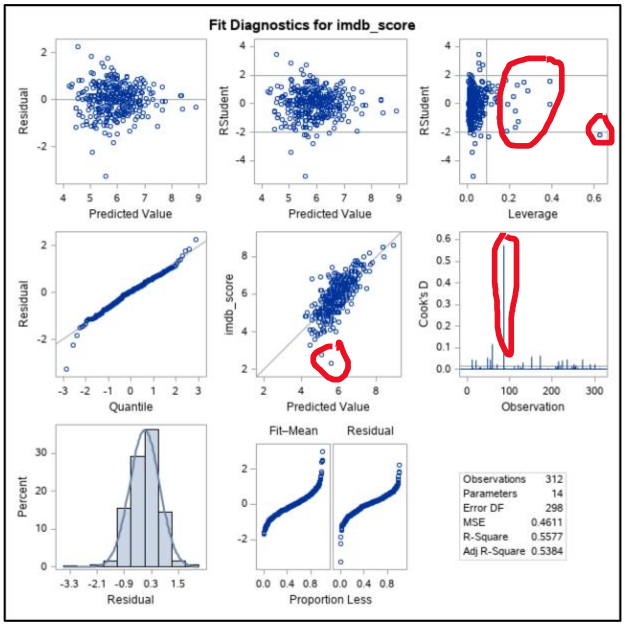
For every 1% change in x	num_critics	gross	users_voted	budget	num_critic & duration	gross & user_voted	user_voted & user_reviewed
imdb score change	-8.93	-1.88	2.69	-2.71	2.4	0.07	0.44



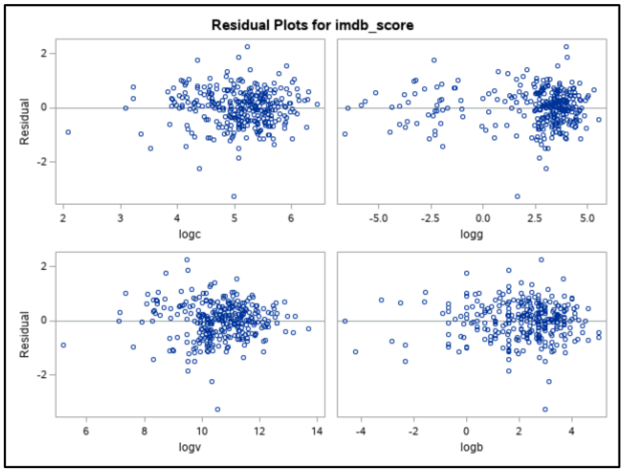
For every 1% change in x	num_critic & user_voted	duration & gross	duration & user_voted	duration & user_reviewed	duration & budget	gross & user_reviewed	Intercept
imdb score change	-0.2	0.48	-0.75	-0.98	0.56	-0.22	13.48

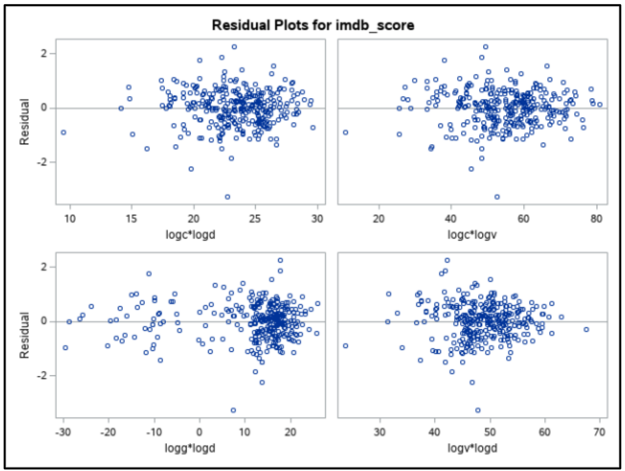
The most influential predictor is number of critics. Holding all else constant, each 1% increase in the number of critics reviewed, imdb scores will fall by 8.93. This is very impactful because the max imdb score is 10. The intercept here isn't very useful because each imdb record will not have missing values for required variables, and the-numbers.com have very extensive data as well for each movie, there will not be a chance where all of these variables equal to 0. The most positively correlated variable is users voted on imdb. This is to be expected, as there's a higher chance of improving scores when more users are voting.

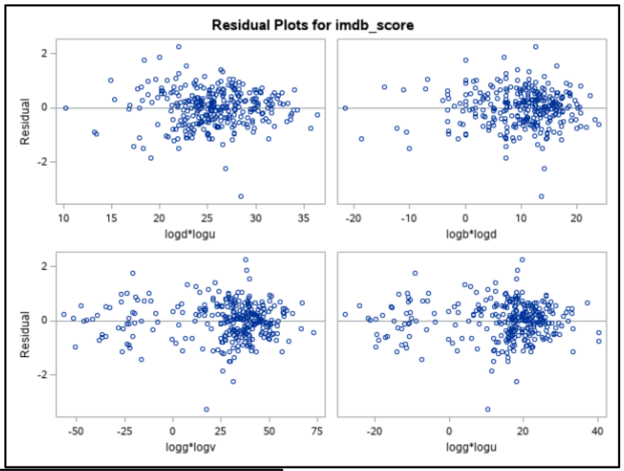
Next, we will confirm that this model meets all the assumptions of linear regression.

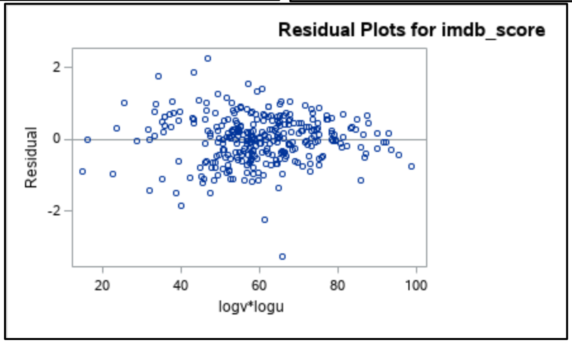


```
Observations     312
Parameters        14
Error DF         298
MSE           0.4611
R-Square      0.5577
Adj R-Square  0.5384
```









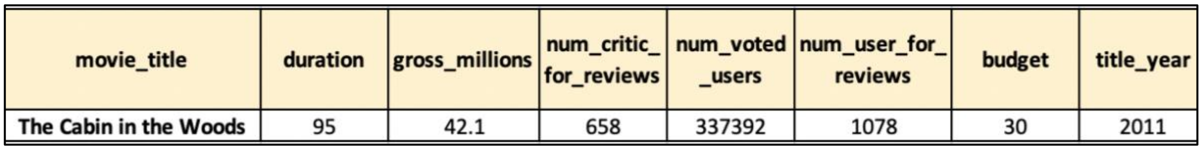
As we can tell from these plots, normality, linearity, and homoscedasticity are all met because of the random, non-pattern, patterns of the residuals distribution. The outliers and high leverage points are circled in red, however they don’t look severe because for a dataset of 312 observations, we can see from the plots that there aren’t many.

Because we introduced interaction terms which guarantees high correlation between the variables in interaction, we are not going to test out multicollinearity here using linear regression. Rridge regression is an alternative to find collinearity for interaction terms, but for the purpose of linear regression analysis it will not be discussed.

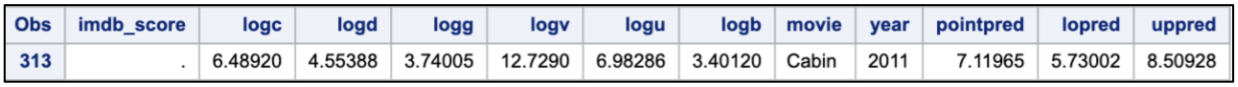
INDIVIDUAL RESPONSE PREDICTION INTERVALS

Below are two horror movies found on Wikipedia, the-numbers.com and imdb.com after the 2000s used to test out how accurate our model is. Here’s a list of data sources:

1. The Cabin in the Woods (2011)



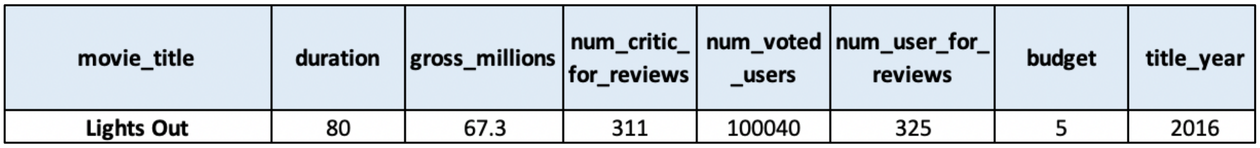
movie_title	duration	gross_millions	num_critic_for_reviews	num_voted_users	num_user_for_reviews	budget	title_year
The Cabin in the Woods	95	42.1	658	337392	1078	30	2011



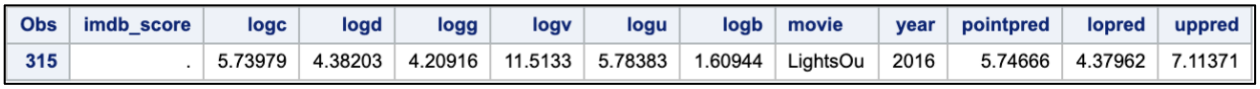
Obs	imdb_score	logc	logd	logg	logv	logu	logb	movie	year	pointpred	lopred	uppred
313	.	6.48920	4.55388	3.74005	12.7290	6.98286	3.40120	Cabin	2011	7.11965	5.73002	8.50928

Our model predicts that "The Cabin in the Woods" should be rated at 7.11965, and the prediction interval is between 5.73002 to 8.50928. The actual score found on imdb.com is 7.0, which proves that our model is pretty accurate.

2. Lights Out (2016)



movie_title	duration	gross_millions	num_critic_for_reviews	num_voted_users	num_user_for_reviews	budget	title_year
Lights Out	80	67.3	311	100040	325	5	2016



Obs	imdb_score	logc	logd	logg	logv	logu	logb	movie	year	pointpred	lopred	uppred
315	.	5.73979	4.38203	4.20916	11.5133	5.78383	1.60944	LightsOu	2016	5.74666	4.37962	7.11371

The point prediction for Lights Out is 5.74666, and the 95% prediction interval is between 4.37962 and 7.11371. The actual score is 6.3, which isn't very far from the point prediction, and inside the prediction interval.

CONCLUSION

Can we make a headline tomorrow such as "People don't trust critic reviews! The more frequently reviewed horror movie the worse the quality it is!" or "Production budget all goes to waste! Horror film producers should spend minimum budget on production, because the more investment, the worse the IMDB ratings!"? Probably not. Since there are so many predictors in our model, to compare two movies that have exactly the same values for all predictors but one, thus holding all other effects constant, is extremely difficult in real life.

Also, in real life, audiences consider more than just these predictors. They also care about who are the hot buzz stars in the cast, the theme/idea of the main plot, CGI technology level, whether plot logic makes sense…etc. And all of those potential predictors subject to personal preferences and thus hard to quantify.

We could also further study our model building process in the future by eliminating the outliers and high leverage data points. This was not done in the current analysis because before, and other the log transformation, the number of outliers in both diagnostics reports don't seem severe, considering this is a large enough sample dataset.

What we are able to learn from this analysis is that our regression model estimates a horror movie's imdb scores, given the predictors we used, rather accurately, at 5% chance of error. We can look for upcoming horror movies of 2019 and beyond and apply this model to get approximation of what others on imdb thinks about the same movies. For entertainment purpose, we can also use this model as a rough personal guideline to evaluate whether we want to watch a movie or not, if we've been hearing buzz about one that we are unsure of.