IDS 702: Final Report

Erika Fox

Summary

The following report uses data from the movie database and reviewing platform, IMDb, to explore the main question: Do movies with higher budgets tend to be better received? We used a proportional odds model to explore this relationship between movie budget and audience reception. We found that budget and audience reception have an unexpectedly inverse relationship that is similar across all movie genres in our analysis, except for horror. However, this relationship is complex and this analysis has its limitations.

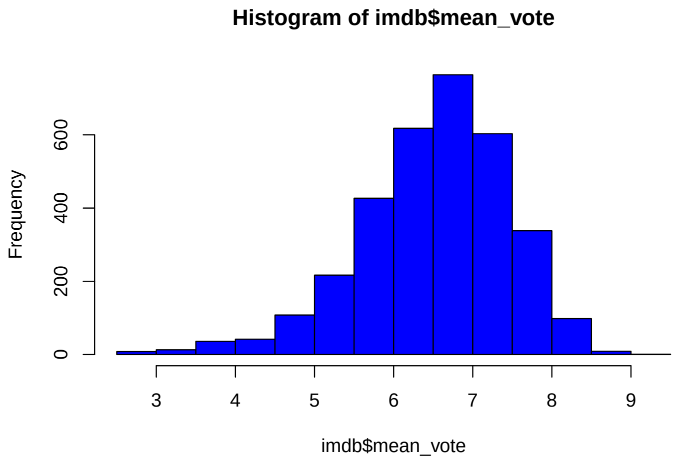
Introduction

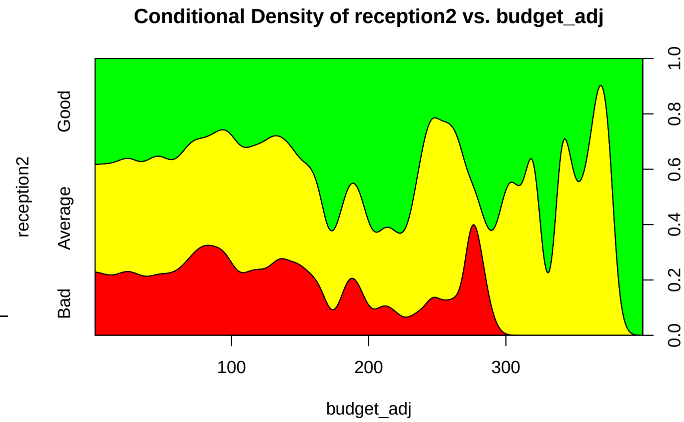
Rather than prioritizing the most enjoyable viewing experience, the U.S. movie market is known for focusing on maximizing gross income. The obvious way production companies go about this mission is to use a high budget, so they can hire A-list actors or boost their project's advertising. Although these efforts may successfully create the illusion of a better-quality movie for audiences, they don't necessarily contribute how well the project is ultimately received. For this reason, we will dissect the relationship between movie budgets and audience response in order to determine once and for all if increasing a movie's budget actually tends to make it to be better received. Questions of interest: Do movies with higher budgets tend to be better received? Is there any evidence that the association between a movie's budget and mean vote score differ by the movie's … Genre? Duration? Cast? Are there any other interesting associations that are worth mentioning?

Data Overview and EDA

The data used in this analysis came from the "IMDb movies extensive dataset" from Kaggle. From this Kaggle posting, we took two csv files, one that focused primarily on baseline movie information (i.e. title, duration, release date, etc.) and another that focused on how movies performed on the IMDb platform (i.e. total number of votes, average vote, etc.). These two files were easily merged using the *imdb_title_id* (a unique code every IMDb movie is given) variable, resulting in one big dataframe with complete information for over 81,000 movies. For this analysis, we chose to only focus on one movie market, movies from the U.S. Additionally, we only wanted to look at movies that had at least 50 reviews from users ($reviews\_from\_users > 50$) on IMDb, as this would prevent us from having skewed results (often movies would have low ratings, but only based on 10 opinions, or vice versa). These cuts to the data brought us down to a total of 3,282 observations (movies). From here, a few format changes were made. *budget*, representing a movies budget, became *budget_adj* (our main predictor variable), by converting this variable from dollars to millions of dollars and adjusting for inflation using an additional Consumer Price Index dataset. Originally, our *genre* variable contained a list of genres for each individual movie. If we were going to have any hope of learning anything from genre, we needed to cut this down so each movie was only matched to one genre. We did this using the following hierarchy: Family, Comedy, Action, Horror, Romance, Drama, Thriller, then other. In other words, if a movie had "Family" in it's genre movie, it would always be classified as a Family movie in our new genre variable, genre2. If not Family, the movie would get classified as the next genre it has in the hierarchy. Finally, we also created the variables *decade*, a factor variable to simplify *release_date* and indicate which decade a movie came from, and *oscar*, a factor variable indicating whether or not a movie had a cast member with an Oscar nomination at the time of the movie's release. To create *oscar*, we utilized an official Academy Awards dataset and cross-referenced it with the *actors* variable in our original data, by *release_date*.

Finally, our response variable for this study is $reception2$, which is an ordered factor variable based on the original variable *mean_vote*. On IMDb, users and critics can go in and rate movies on a scale from 1-10, and *mean_vote* represents the average of all ratings given on a particular movie (including reviews from both regular users and critics). The histogram on the right below shows how this average is distributed, and helped us create a categorical measure for how movies are received, where movies with a *mean_vote* of 7 and higher are considered "Good" or well received, movies with a $mean\_vote$ between 6 and 7 are considered “Average”, and movies with a $mean\_vote$ of less than 6 are considered “Bad”. The plot on the right below describes how the conditional distribution of $reception2$ changes in relationship to $budget\_adj$.





Model

Our final model for this analysis is a proportional odds model represented by the following formula:

$$log(\frac{P(y_i \leq j|x_i)}{P(y_i > j|x_i)}) = \beta_{0j} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, j = 1, \ldots, J-1$$

where y_i is the categorical variable reception2, with “Good”, “Bad” and “Average”, x_i includes the predictors: $budget\_adj^3$, $log(reviews\_from\_users)$, $genre2$, $duration$, and $reviews\_from\_critics$, and β_i is a vector representing the predictor coefficients.

Selection: In order to arrive at this final model, we used AIC stepwise. For the null model input, we used the simplest possible model that would help us reach our research goals, with only had $budget\_adj$ as a predictor for our response variable, $reception2$. For the full model input, we used all feasible* predictors, including: $budget\_adj$, $oscar2$, $genre2$, $reviews\_from\_users$, $reviews\_from\_critics$, and $duration$.

* some potential predictors needed to be excluded, either because they were too complex (i.e. $production\_company$), or because of co-linearity concerns (usa_gross_income)

Note that we also would have included interactions here in the full model, if we were to find any that seemed concerning during EDA. The only interaction we found of interest was $genre2 : budget\_adj$. However, we ultimately decided to exclude it anyway, as the way we modified $genre$ for this analysis already feels like enough of a stretch. A few anova tests also helped confirm this is the best model to move forward with.

Transformations: AIC stepwise returned what became our final model, except the variable $reviews\_from\_users$ was not yet logged, and $budget\_adj$ did not yet have the polynomial transformation. The decision to add these transformations happened during model assessment, when we found that the binned residuals plot for these variables were not satisfactory. See appendix for plot comparisons.

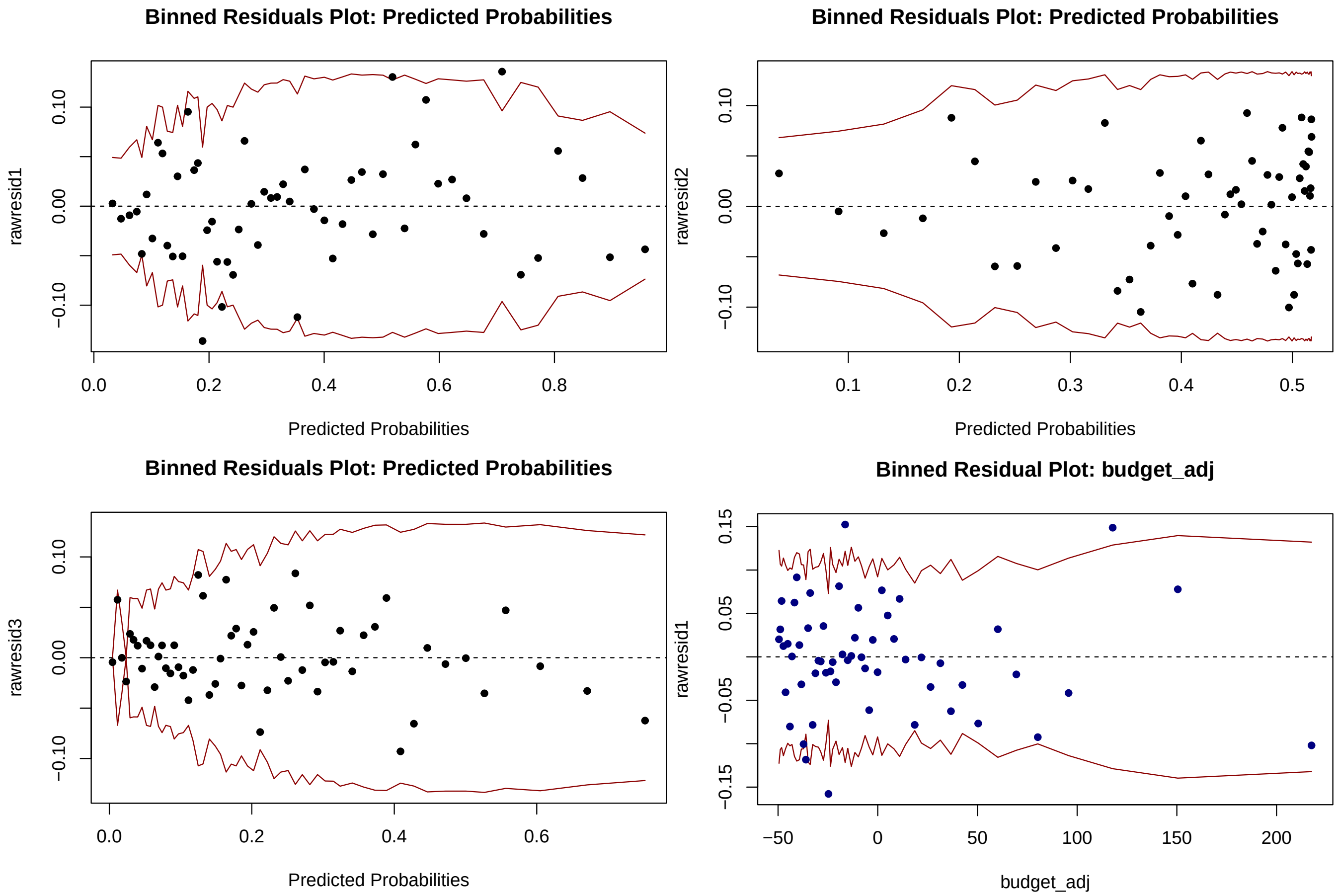
Assessment: On the previous page there is a summary for our model (Table 1), that includes confidence intervals for each of our variables. This summary shows that none of our variables' confidence intervals contain zero, which means that we successfully built a model with significant predictors.Our model satisfactorily meets the assumptions for proportional odds models. Evaluating multicollinearity for this model was not simple. Ideally we want our GVIF scores to all stay below 5, but because this model has a polynomial transformation on it, things got put out of whack. The adjusted VIF scores ($GVIF^(1/(2 * Df))$) are 2.12, 8.42, 1.24, 1.11 and 1.42 for $poly(budget\_adj, 3)$, $log(reviews\_from\_users)$, $genre2$, $duration$, and $reviews\_from\_critics$, respectively. Even though this is not a perfect list of scores, because the values were in range before adding the polynomial transformation, the values are still not astronomically big, and there is frankly not much we can do to fix this, we settled on considering multicollinearity to be satisfied.

	Model 1
poly(budget_adj, 3)1	32.20^{*}
	$[32.16; 32.25]$
poly(budget_adj, 3)2	-12.80^{*}
	$[-12.83; -12.78]$
poly(budget_adj, 3)3	12.11^{*}
	$[12.08; 12.13]$
logreviews	-0.94^{*}
	$[-1.06; -0.82]$
genre2Comedy	-0.32^{*}
	$[-0.53; -0.12]$
genre2Drama	-1.13^{*}
	$[-1.39; -0.87]$
genre2Family	-0.63^{*}
	$[-0.96; -0.30]$
genre2Horror	1.44^{*}
	$[1.17; 1.72]$
genre2Other	-1.13^{*}
	$[-2.25; -0.02]$
genre2Romance	-0.73^{*}
	$[-1.07; -0.39]$
genre2Thriller	-0.52^{*}
	$[-0.84; -0.19]$
duration	-0.03^{*}
	$[-0.04; -0.03]$
reviews_from_critics	-0.00^{*}
	$[-0.00; -0.00]$
Good|Average	-9.48^{*}
	$[-10.22; -8.74]$
Average|Bad	-7.19^{*}
	$[-7.90; -6.47]$
AIC	5956.95
BIC	6048.39
Log Likelihood	-2963.47
Deviance	5926.95
Num. obs.	3282

* 0 outside the confidence interval.

Table 1: Final Model Summary

The first three out of the four plots below, labeled "Binned Residuals Plot: Predicted Probabilities", allow us to assess the rest of our assumptions using three separate groups of raw residuals, representing each of the three classes from our response variable, $reception2$. For all three plots, we can see that 95% of points are within the bands, and there does not seem to be a trend to be concerned about. The fourth plot here, labeled "Binned Residual Plot: budget_adj" assesses our main predictor, *budget_adj* alongside the first group of raw residuals, once we added the cubic polynomial transformation. Just like the previous plot, we can see that 95% of points are within the bands, and there does not seem to be a trend to be concerned about.



Finally, the following table contains additional metrics to help us understand how our model is performing.

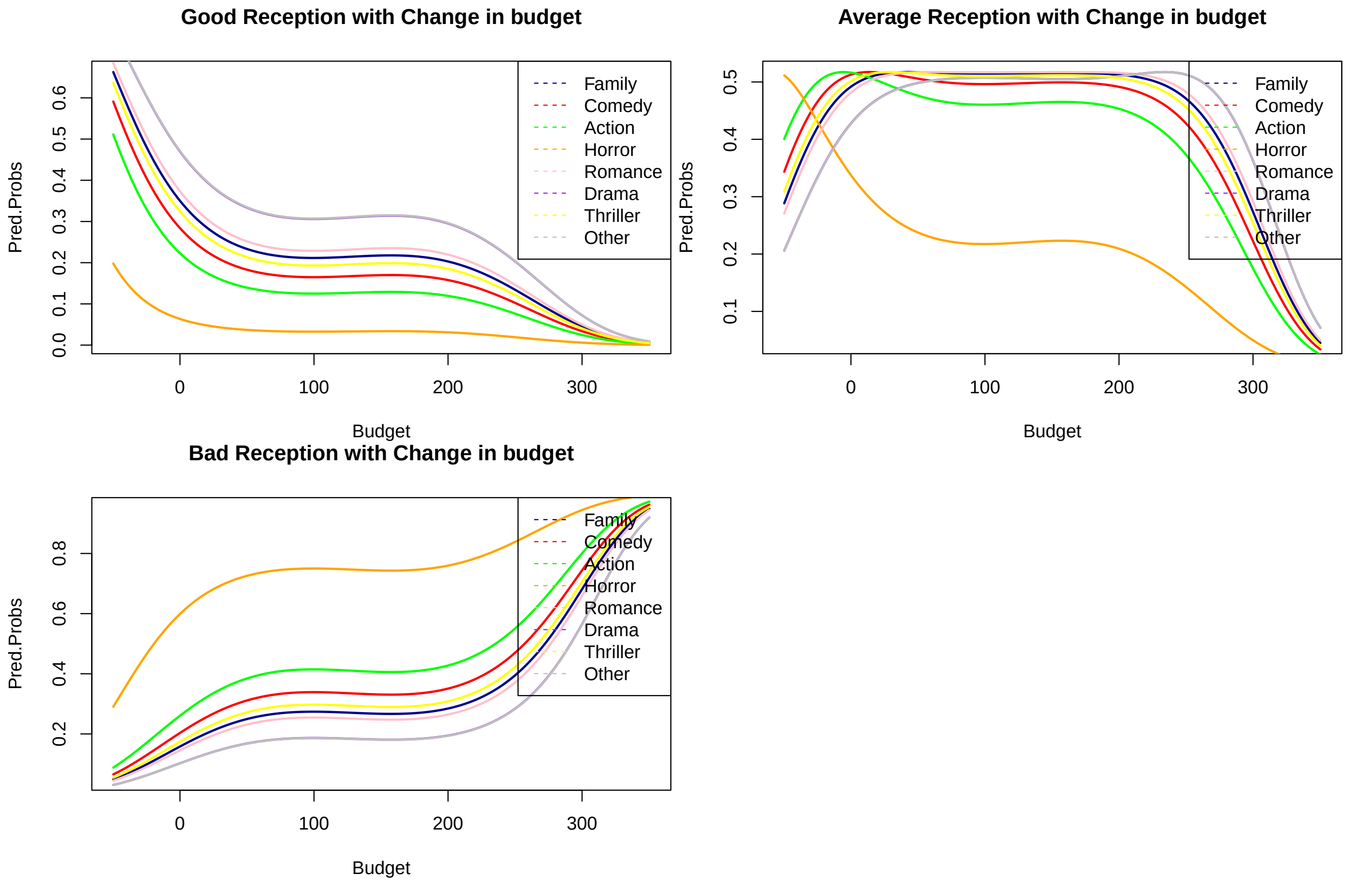
	Sensitivity	Specificity	Pos Pred Val.	Neg Pred Val.	Precision	Recall	F1	Bal. Acc.
Good	0.61	0.83	0.67	0.79	0.67	0.61	0.64	0.72
Average	0.66	0.55	0.50	0.70	0.50	0.66	0.57	0.61
Bad	0.33	0.92	0.56	0.82	0.56	0.33	0.41	0.62

This table shows a bit of variety in terms of the main performance measures, but across the board, we have decent accuracy scores of 0.72 for the "Good" class, 0.62 for the "Average" class, and 0.62 for the "Bad" class.

When it comes to specificity, we have scores above 0.50 across all three classes, getting as high as 0.92 for the "Bad" class. However, on the other hand, sensitivity is low for the "Bad" class at 0.33. This means that our model has a hard time classifying movies as "Bad" when they are actually poorly received, but does a good job classifying movies as something other than "Bad" when they are in fact, well or averagely received.

Interpretation: Because we have a both a log and a cubic polynomial transformation in our model, the interpretation is quite complicated and would be messy to describe in words. Therefore we must rely on the following graphs. Each of these graphs display how each category in our response variable *reception*2 changes by our main predictor, *budget_adj*. The graphs are stratified by the *genre*2 variable.

These plots definitely demonstrate that movie budget and audience reception have a complicated relationship, and there are some unexpected insights to discuss. The graphs show more extreme behavior at the extreme ends of the range of *budget_adj*, whereas the trend appears to be relatively flat in the middle for all three reception classes. Most of the genres appear to share similar trends, despite differing in the actual values of predicted probabilities. Horror is really the only genre that diverges from the group, particularly in the "Average" class.



Interestingly, budget_adj and reception appear to have an inverse relationship, where our highest budget movies have the best probability of classifying as "Bad", whereas our lowest budget movies appear to have the best probability of classifying as "Good", for most genres. As mentioned before, the trend of predicted probabilities tends to flatten in the middle for all three classes, but this flat trend line sits the highest for the "Average" class at around 0.5. In other words, according to this interpretation, if a production were to go with a middle-of-the-road budget of between about 100 and 300 million dollars, that movie would have about a 50% chance of getting Average reception for most genres.

Conclusions:

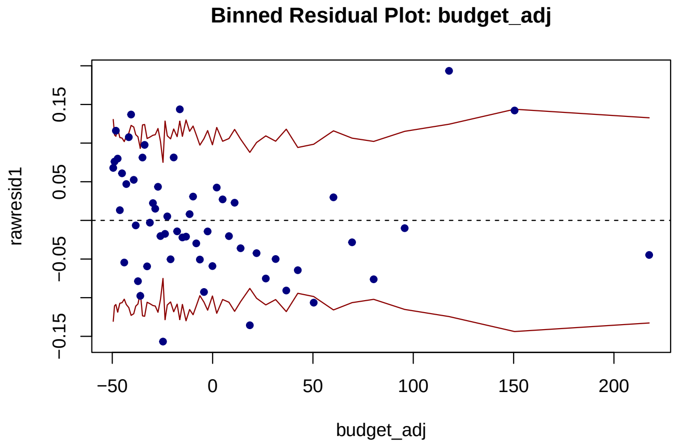
In this analysis, we sought out to examine the impact a movie's budget has on how that movie is ultimately received. The proportional odds model we built indicates that there *is* a relationship between these two variables, our response: *reception*2 (an ordered factor, Good, Average, and Bad, indicating how well the movie was received) and our main predictor: *budget_adj* (movie budget in millions of dollars, adjusted for inflation). According to our model, movies with higher budgets do not generally tend to be better received. Our model not only considered *budget_adj* to be a significant predictor of *reception*2, but also considered the rest of the predictors we ended up using to be significant as well. Using interpretive plots, we were able to find evidence that the trends in predicted probability for each reception class are generally similar across genre, although Horror movies seem to operate a bit differently.

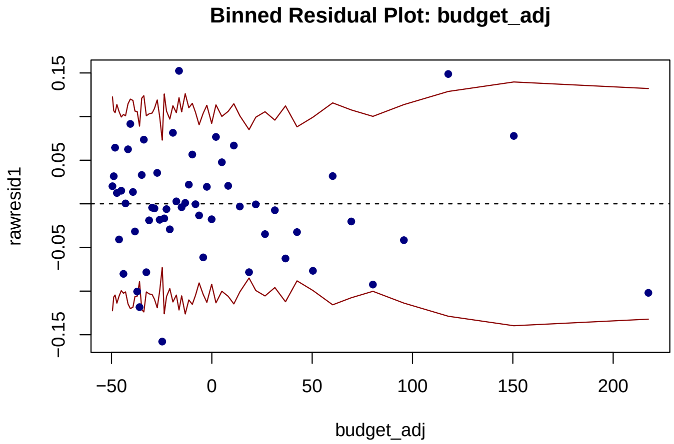
Limitations:

- The genre variable needed to be heavily modified in order for us to learn anything from it. In this process, we lost a lot of information that could have altered some findings from this analysis.
- There are many factors that go into whether or not an audience member receives a movie well that are beyond the scope of our data, for example we have no measure of emotion, how well certain jokes landed, etc.
- Our model's accuracy, sensitivity and specificity measures leave room for improvement. Particularly, our model has high specificity when it comes to classifying movies as "Bad" or poorly received, meaning that we might be missing out on some insightful findings as to what makes a movie poorly received.

Appendix

Left, before transformations. Right, after transformations.





Left, before transformations. Right, after transformations.