Which factors affect the odds that a movie passes the Bechdel test?

The Kable Guys: Ryan Mitchell, Jerry Hou, Arjun Prabhakar, Nathan Huang

December 5, 2021

1 Introduction and Data

The film industry is home to rampant gender inequality, evidenced by a disproportionately low amount of female directors, stars, and producers. This imbalance manifests itself into the films it invades, leading to a widespread lack of women on the silver screen. To further analyze this issue, we are investigating the representation of women in films based on a measure referred to as the *Bechdel test*. For a film to pass the Bechdel test, it must have at least two female actresses who talk to each other about a topic besides a male character (Garber 2015). This simple test, coined by cartoonist Alison Bechdel in a 1985 comic, helps us gauge the level of gender equality in a cinematic work (Garber 2015). Unsurprisingly, many blockbuster films fail the Bechdel test due to an infamous lack of diversity in the film world.

The data used for this analysis is a part of the Tidy Tuesday repository, a collection of datasets released to the general public in an effort to produce varied data visualizations. The dataset comes from FiveThirtyEight and compiled statistics from BechdelTest.com and The-Numbers.com (Hickey 2014). FiveThirtyEight used the former to determine if films passed the Bechdel test and the latter to obtain financial information on the films. The dataset has 1794 observations, each representing a film released from the years 1970 to 2013, and 34 variables, including genre, IMDB rating, domestic gross, international gross, budget, awards, and whether or not the film passes the Bechdel test.

Since our dataset includes rating and grossing data for films, we want to take a closer look at which factors relate to the odds that a movie passes the Bechdel test. A New York Times analysis of scores from Metacritic indicated that films directed by women of color were the best received by critics (Buckley 2020). Additionally, several of the best-grossing movies from 2020 had predominantly female casts, so an analysis of a greater number of films can give us a better picture of the relationship between metrics about a film and its odds of Bechdel Test passage.

The general research question that we wish to explore is: Which factors (including critical rating, advisory rating (i.e. PG-13 vs. R), audience international grossing) affect the odds that a movie passes the Bechdel test? We hypothesize that movies that are newer, have a larger domestic grossing, are better rated by critics and audiences, and are rated for younger audiences have greater odds of passing the Bechdel test.

To approach this research question, we will examine the log odds of a film passing the Bechdel test as our response variable. Since this is a binary variable, we will create a logistic regression model that can predict probabilities of passing for each film.

1.1 Dataset Transformations

Before conducting exploratory data analysis, we implemented some transformations on the dataset. We removed all observations that did not have values for any of the initial predicted variables outlined in Section 2. The absence of data for all of these variables was not strongly correlated with probability of passing the Bechdel test; of the films removed, approximately 54% failed the Bechdel test and 46% passed (and in the full dataset, approximately 55% failed the Bechdel test and 45% passed.)

The original dataset contained films with 12 distinct advisory ratings. We condensed these into four categories: child audiences (G/PG/TV-PG), teenage audiences (PG-13/TV-14), adult audiences (R/X/NC-17), and unrated. In order to increase the usefulness of conclusions, we decided to scale monetary variables, including budget, normalized domestic gross, and normalized international gross, to millions of dollars. Additionally, budget and gross values used in analysis were all normalized to 2013.

1.2 Exporatory Data Analysis

1.2.1 Domestic Gross

We hypothesized that films with a higher domestic gross will have higher odds of passing the Bechdel test. The distribution of domestic gross can be seen below:

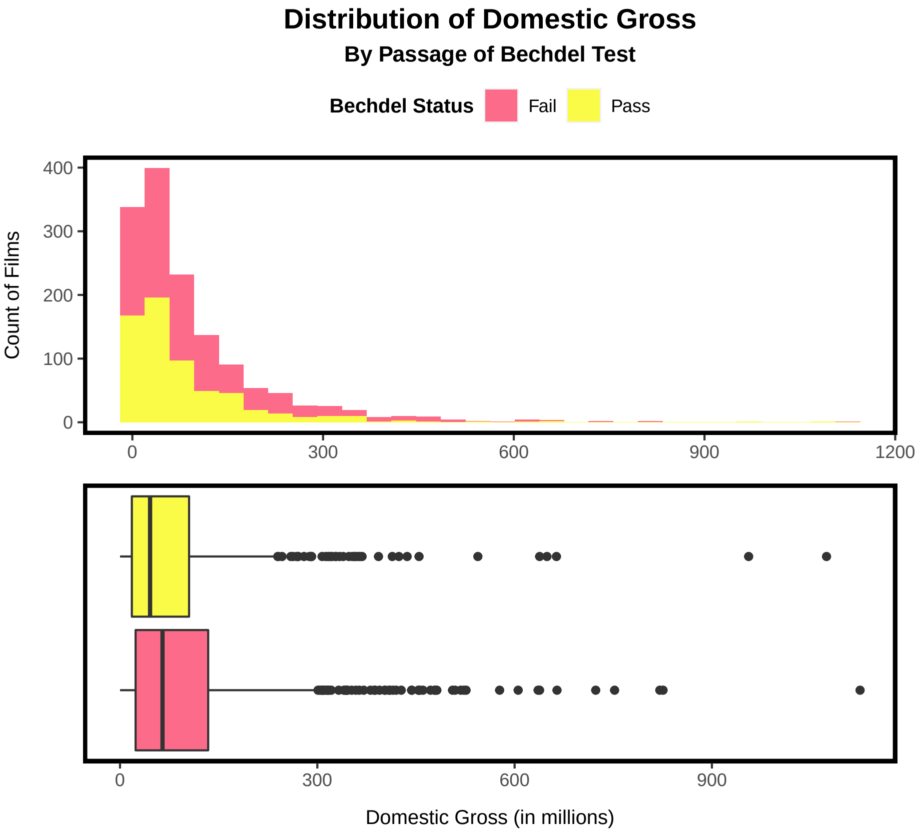


Table 1: Summary of Domestic Gross for Passing Films

Mean	Median	Std. Dev	IQR
80.939	45.7	107.791	87.049

Table 2: Summary of Domestic Gross for Failing Films

Mean	Median	Std. Dev	IQR
104.783	64.6	126.85	110.391

Overall, the distributions of domestic gross for passing and failing films are both heavily skewed right. Using the median, the distribution of failing films has a noticeably higher center ($64.6m v. $45.7m) and larger spread ($110.4m v. $87m), as well as higher first and third quartile values. This initial data analysis goes against our hypothesis.

1.2.2 Critical Rating

We hypothesized that films that are higher rated by critics and audiences will have higher odds of passing the Bechdel test. The distribution of IMDb ratings can be seen below:

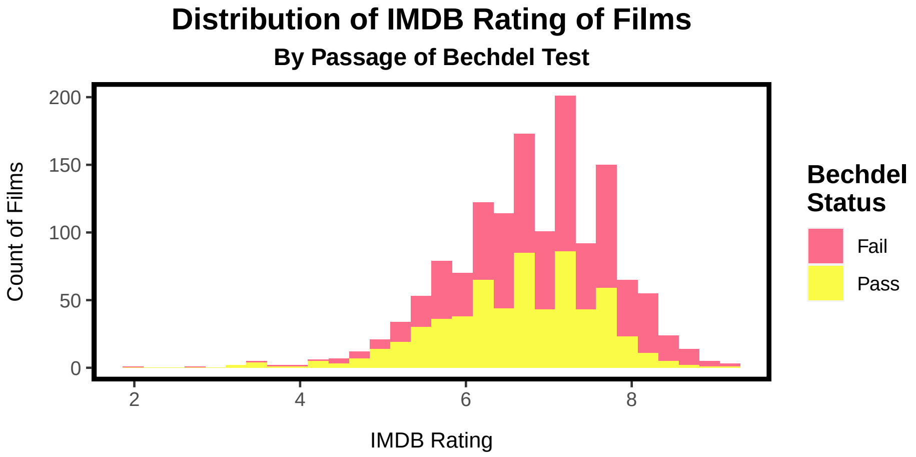


Table 3: Summary of IMDB Ratings for Passing Films

Mean	Median	Std. Dev	IQR
6.624	6.7	0.947	1.3

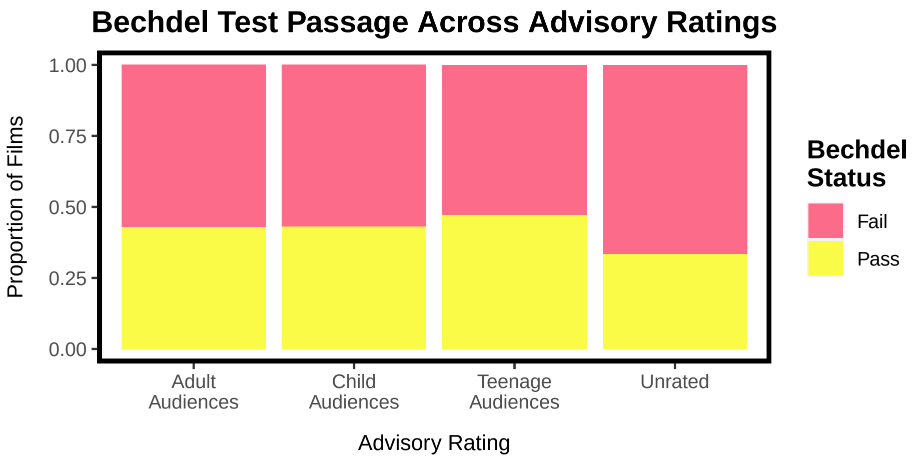
Table 4: Summary of IMDB Ratings for Failing Films

Mean	Median	Std. Dev	IQR
6.922	7	0.943	1.2

The distributions of IMDB ratings for passing and failing films are roughly equivalent, both having a slight left skew. The distribution of failing films has a slightly higher center (7 v. 6.7) and a slightly smaller spread (1.2 v. 1.3). Unlike what we hypothesized, the histogram does not display significant evidence that films with higher critical and audience ratings have higher odds of passing the Bechdel test.

1.2.3 Advisory Rating

We hypothesized that films rated for younger audiences will have higher odds of passing the Bechdel test. The rate of Bechdel test passage for different advisory ratings can be seen below:



The bar graph shows that the percentage of films passing the Bechdel test is roughly equivalent amongst all advisory rating groups, but there is a noticeably lower proportion of passing films amongst unrated films. Films rated for teenage audiences have a slightly higher proportionate of passes than those rated for adult and child audiences. Unlike what we hypothesized, the histogram does not display significant evidence that films rated for younger audiences have higher odds of passing the Bechdel test.

2 Methodology

For our analysis, we used the binary pass/fail status of whether or not a film passes the Bechdel Test as our response variable. Since this is a categorical variable with two levels, and we are interested in predicting the probability a film passes the Bechdel Test, we used a logistic regression model.

Initially, we considered the following variables in the dataset to be predictors:

- `year` - Year of the film's release
- `budget_2013` - Budget of the film normalized to the year 2013
- `domgross_2013` - Domestic gross of the film normalized to the year 2013
- `intgross_2013` - International gross of the film normalized to the year 2013
- `rated` - Content rating assigned to the film
- `metascore` - Metacritic rating of the film, from 0-100
- `imdb_rating` - IMDB rating of the film, from 0-10
- `runtime` - Runtime of the film, in minutes

2.1 Variable Transformations

Before creating the initial regression model, we transformed multiple predictor variables in the dataset to increase the usefulness of conclusions. The variable `year` will be redefined as `years_1970`, a numerical metric representing how many years after 1970 that the film was released. Budget, international gross, and international gross will all be mean-centered to increase the practicality of the intercept.

2.2 Model Selection

We fit a model with the predictors outlined above and used both forwards and backwards selection to obtain the model with the lowest AIC.

Table 5: Resulting Model AIC Removing Each Predictor Variable

Model Options	AIC
Full Model	1845.3
- Years Since 1970	1845.9
- Domestic Gross	1846.9
- Runtime	1848.7
- International Gross	1850.1
- Advisory Rating	1851.5
- Metacritic Rating	1857.3
- IMDB Rating	1890.0
- Budget	1897.7

Both selection methods resulted in the same model, removing none of the 8 initial predictor variables. As shown in the table, the full model is the one producing the lowest AIC value.

We then examined the interactions between budget and domestic/international gross. This required a drop-in-deviance test with the following hypotheses:

$$H_0 : \beta_{\text{Centered Budget : Centered Domestic Gross}} = \beta_{\text{Centered Budget : Centered International Gross}} = 0$$

$$H_a : \text{ At least one new coefficient } \beta_j \text{ is not equal to 0.}$$

The results of the drop-in-deviance test were as follows:

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1403	1823.310	NA	NA	NA
1401	1818.319	2	4.991125	0.08245

The p-value is significant at $\alpha = 0.1$, so we rejected the null hypothesis. The data provided sufficient evidence that at least one of the coefficients associated with the interactions between budget and domestic/international gross were not equal to 0, so we added these terms to our model.

2.3 Multicollinearity

We then checked our model for possible multicollinearity using the variance inflation factor (VIF). If VIF for a predictor variable was greater than 10, we considered that as an indication of concerning multicollinearity.

Table 7: VIF of Predictors Using Selected Model

	VIF
years_1970	1.288701
budget_2013Cent	2.360871
domgross_2013Cent	14.667360
intgross_2013Cent	18.177395
ratedChild Audiences	1.450409
ratedTeenage Audiences	1.387362
ratedUnrated	1.036882
metascore	2.342647
imdb_rating	2.607089
runtime	1.517951

	VIF
budget__2013Cent:domgross__2013Cent	16.886853
budget__2013Cent:intgross__2013Cent	18.212259

This revealed that domestic gross and international gross were highly correlated, along with their interactions with budget, so we considered two alternative models to eliminate this multicollinearity: one without domestic gross and its interaction with budget, and one without international gross and its interaction with budget.

Table 8: AICs of Possible Multicollinearity-Removing Models

Model Option	AIC
Removing Domestic Gross	1845.564
Removing International Gross	1844.717

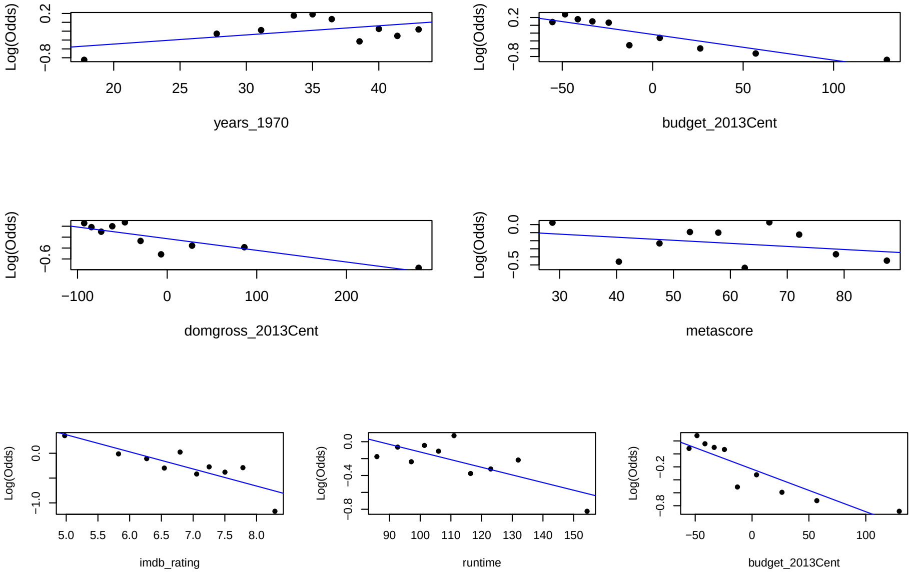
The model removing international gross and the interaction between international gross and budget produces a lower AIC, so we will proceed using this model. A second multicollinearity check using this model reveals no VIF values over 10.

2.4 Model Conditions

Using our new model absent of any multicollinearity, we then checked for the proper conditions of a logistic regression model, namely linearity, randomness, and independence.

2.4.1 Linearity

We checked for the linearity of our model using empirical logit plots. We divided each quantitative predictor into 10 intervals, calculating the mean value of the predictor in each interval. We then computed the empirical logits and plotted them versus the mean value of the predictor in each interval, as seen below:



Based on the empirical logit plots, the linearity condition is satisfied. There appears to be a linear relationship between the log odds and each predictor variable.

2.4.2 Randomness & Independence

The randomness condition is satisfied. The films in the dataset can be reasonably treated as a random sample of films released from 1970 to 2013, and there is no indication that the data was obtained from a non-random process.

The independence condition is satisfied. We can reasonably assume that the films in the sample are independent from one another, and that the result for any one film does not impact the result of any others.

3 Results

Our final model was as follows:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.20963	0.62782	1.92673	0.05401	0.17916	2.24554
years_1970	0.01610	0.00839	1.91819	0.05509	0.00236	0.02999
budget_2013Cent	-0.01094	0.00156	-7.01592	0.00000	-0.01354	-0.00841
domgross_2013Cent	0.00030	0.00067	0.44524	0.65615	-0.00082	0.00138
ratedChild Audiences	0.39461	0.18230	2.16467	0.03041	0.09496	0.69502
ratedTeenage Audiences	0.36256	0.13543	2.67707	0.00743	0.14020	0.58588
ratedUnrated	-0.95425	0.51543	-1.85135	0.06412	-1.84955	-0.13142
metascore	0.01703	0.00495	3.44261	0.00058	0.00893	0.02522
imdb_rating	-0.61942	0.09651	-6.41826	0.00000	-0.77973	-0.46209
runtime	0.00852	0.00351	2.43137	0.01504	0.00276	0.01430
budget_2013Cent:domgross_2013Cent	0.00002	0.00001	2.76984	0.00561	0.00001	0.00003

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 1.210 + 0.016\ years\_1970 - 0.011\ budget\_2013Cent + 0.0003\ domgross\_2013Cent + 0.395\ ratedChildAudiences + 0.363\ ratedTeenageAudiences - 0.954\ ratedUnrated + 0.017\ metascore - 0.619\ imdb\_rating + 0.009\ runtime + 0.00002\ budget\_2013Cent : domgross\_2013Cent$

3.1 Model Interpretations

As aforementioned, we were interested in determining which factors impact the odds that a movie passes the Bechdel test. We initially hypothesized that movies that are newer, have a larger domestic gross, are better rated by critics and audiences, and are rated for younger audiences have greater odds of passing the Bechdel test. We will draw conclusions from our model using a significance level of $\alpha = 0.1$ and a 90% confidence interval.

Regarding release date of a film, the model suggests that for each additional year since 1970 that a film was released, we expect the odds of that film passing the Bechdel test to multiply by a factor of 1.016, holding all else constant. The p-value is significant and the confidence interval is entirely positive, suggesting statistically significant evidence that a newer film has larger odds of passing the Bechdel test.

Regarding advisory rating, the model suggests that we expect the odds that a film made for child audiences passes the Bechdel test to be 1.484 times the odds that a film made for adult audiences passes the Bechdel test, holding all else constant. A similar, albeit less pronounced effect exists for films made for teenage audiences. The p-values for both child and teenage audiences are significant, and both confidence intervals are entirely positive, suggesting statistically significant evidence that films rated for younger audiences have larger odds of passing the Bechdel test.

Regarding critical and audience rating, the results initially seem contradictory. Holding all else constant, the effect of Metacritic score is statistically significant and suggests higher ratings increases the odds of a film passing the Bechdel test, but the opposite effect is found for IMDB rating. However, a notable difference between the two metrics is that Metacritic scores are submitted from film critics while IMDB ratings are submitted from general audiences; this may reflect a difference between critics and audiences when it comes to valuing female representation as an indicator of a film's quality.

Contrary to our hypothesis, the model did not provide sufficient evidence that domestic gross impacts the odds a film passes the Bechdel test (p = 0.656). However, the model suggests that for every million dollar increase in a film's budget, we expect the odds of that film passing the Bechdel test to multiply by a factor of 0.989, holding all else constant. With a statistically significant p-value and an entirely negative confidence interval, there is sufficient evidence that films with lower budgets have higher odds of passing the Bechdel test. There also is a small, albeit significant impact of the interaction between budget and domestic gross, suggesting that the impact of a million dollar increase in domestic gross on the odds a film passes the Bechdel test increases as budget increases.

Runtime of a film also impacts the odds of a film passing the Bechdel test, with longer films increasing the odds of passage. Specifically, the model suggests that for every minute increase in runtime, we expect the odds a film passes to multiply by a factor of 1.009, holding all else constant.

Regarding the research question, the model shows us that year, budget, advisory rating, Metacritic rating, IMDB rating, runtime, and the interaction between budget and domestic gross all impact the odds a film passes the Bechdel test.

3.2 Model Fit

After interpreting our model, we examined how well our model was at predicting the data to better understand the validity of our conclusions. We used a threshold of 0.5 to distinguish between predicted fails and predicted passes.

Table 10: Predictions vs Actual Results (0 = Fail, 1 = Pass)

Actual Status	Predicted Fail	Predicted Pass
0	587	199
1	307	321

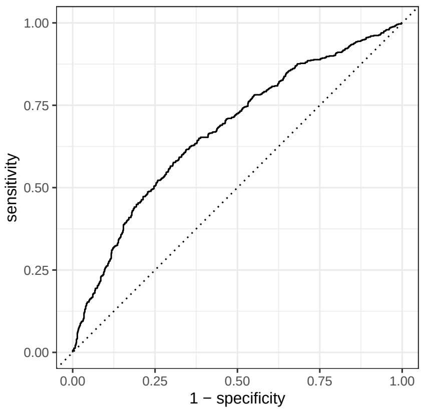
Table 11: Model Accuracy Statistics

Statistic	Value
Sensitivity	0.511
Specificity	0.747
Positive Predictive Value	0.617
Negative Predictive Value	0.657
Misclassification Rate	0.358

The model has a misclassification rate of 35.8%, suggesting it can accurately predict Bechdel test results for about two thirds of films. The model performs much better at correctly predicting failing films, as evidenced by a significantly higher specificity (74.7%) compared to sensitivity (51.1%).

3.3 ROC and AUC

We then wanted to examine the ROC curve to et a better idea of our model's misclassification rate at different thresholds.



Our model has an area under the curve (AUC) value of 0.673, indicating that although the model is not superb, it performs noticeably better than a random classifier (which would have an expected AUC value of 0.5.)

4 Discussion & Conclusion

Our original research question was: **Which factors affect the odds that a movie passes the Bechdel test?** We hypothesized that newer movies, movies designed for younger audiences, and movies with more favorable metrics (grossing/ratings) are more likely to pass the Bechdel test, which is a crude indicator of a film's gender diversity.

We learned that newer movies tend to be more likely to pass the Bechdel test when the other predictors are held constant. This association supports the logical base of our hypothesis, suggesting that over time, the film industry has seen more opportunities for women arise. This reflects both a general societal shift towards further gender equality and the successes of advocacy efforts for greater female representation in film.

We also learned that films made for child and teenage audiences are more likely to pass the Bechdel test. This highlights how children are fortunately exposed to more gender diversity in the world of cinema, but also reflects how more adult themes, like romance, are restrictive in the plotlines they allow for women.

4.1 Our Model

The model used was a logistic regression model as we were working with a binary response variable. Using backward selection, we landed on 8 predictor variables, including the film's release year (in terms of years after 1970), domestic gross, international gross, IMDB rating, Metacritic rating, budget, runtime, and advisory rating; then, a drop-in-deviance test prompted the inclusion of the interaction between budget and domestic/international gross. Analyzing the variance inflation factor (VIF) of our predictors, we found that domestic and international gross (and their interactions with budget) were highly correlated, leading to the removal of metrics related to international gross from our model.

Our model had a sensitivity of 51.1% and specificity of 74.7%, implying that it was much better at identifying movies that did not pass the Bechdel test. Examining our AUC, we found a value of 0.673, showing that our classifier performed better than one that would categorize films at random. Furthermore, all model conditions (linearity, randomness and independence) were satisfied. Overall, we constructed a logistic regression model with robust assumptions and a classification process that, while not excellent, performed better than sorting by random chance.

4.2 Model Limitations and Improvements

We also had parts of our hypothesis that our model did not validate. Perhaps the most surprising was the observed relationship between a film's critical rating and its likelihood of Bechdel test passage: our model indicated that for every one point increase in a film's IMDB rating, we expect the odds it passes the Bechdel test to multiply by a factor of 0.538, holding all else constant. This suggests that general audiences may not view female representation as a true indicator of a film's quality.

A significant limitation of our analysis is that the esoteric nature of the Bechdel test probably contributes to this outcome, as the test criteria about female characters speaking to each other about something besides a male character might not relate much to the metrics of a film. Additionally, the analysis could possibly be improved if we used a gender equality rating or score as our response variable instead of the binary Bechdel test variable. This would allow for a multiple linear regression model that would add more nuance to the level of interaction between female characters.

Furthermore, our model proved not to be a perfect method of prediction. It incorrectly classified over a third of films, and had particular difficulty predicting passing films. Regardless, the model was still evidently useful, performing substantially better than a random classifier would.

4.3 Future Work

Future work should look more broadly at the diversity of a movie and possibly conduct an analysis with some kind of a diversity score as a predictor variable for success metrics of a movie (such as critical reception). This analysis could take a more holistic look at representation in the film industry and provide us with further insight about whether diverse movies tend to be more successful. Analysis could also be stratified for different decades, as it might be revealing to see which factors were useful for predicting Bechdel test passage in different historical eras. Being that this model only considered interactions between budget and gross, further research could also examine more of these terms (such as the interaction between year and rating, to, for example, examine how female representation in films made for adult audiences has changed over time.) It would also be useful to understand how the genre of films impacts Bechdel test passage. Finally, a similar analysis could be conducted for television shows, as this medium requires audiences to keep track of characters for much longer periods of time.

5 References

Buckley, Cara. 2020. "More Women Than Ever Are Directing Major Films, Study Says." *The New York Times*. The New York Times. https://www.nytimes.com/2020/01/02/movies/women-directors-hollywood.html.

Garber, Megan. 2015. "Call It the 'Bechdel-Wallace Test'." *The Atlantic*. Atlantic Media Company. https://www.theatlantic.com/entertainment/archive/2015/08/call-it-the-bechdel-wallace-test/402259/.

Hickey, Walt. 2014. "The Dollar-and-Cents Case Against Hollywood's Exclusion of Women." *FiveThirtyEight*. ABC News Internet Ventures. https://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/.