Stat 210 Project

Sofia Silvosa

Introduction

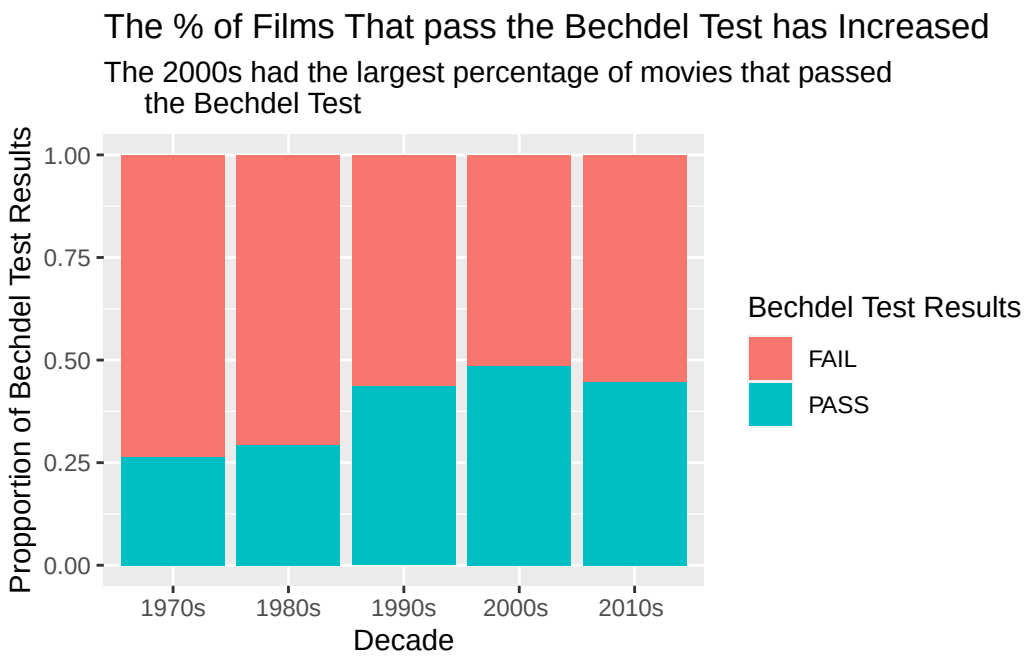
Background and Significance: In the entertainment sphere,there has been recent discussion on adequate female representation in film. Many in the film industry have recently praised the increased representation of women in film (Nolfi,2017).For a while Hollywood produced very little movies with female protagonists; let alone movies with appropriate representation of women (Goodman, 2017). However, merely having female protagonists does not paint the full picture if whether a movie actually does a good job in female representation. Many have used the so-called Bechdel test to make this judgement. This test was developed by Allison Bechdel in 1985 and has recently become a digital sensation (Hickey, 2014).The Bechdel test is a simple test which deems that a movie has adequate female character representation in their film if it passes the following three criteria:(1) it has to have at least two women in it, who (2) who talk to each other, about (3) something besides a man (https://bechdeltest.com/). Many have argued that there has been significant increase of female representation throughout the decades thanks to increased diversity initiatives and more women at the helm of the film industry (UCLA-Hollywood-Diversity-Report-2022-Film).

Research Question & Hypothesis: We're interested in evaluating what variables are important in predicting whether a movie passes or fails the Bechdel test. We are interested, specifically, if the time period in which a movie was released predicts whether a movie passes the Bechdel test. In other words, does the decade a movie was released predict whether it passes the Bechdel test while controlling for other pivotal variables in our model? As shown by recent news and media, more modern movies seem to have more increased representation of women in their films. Thus, we predict that the time period in which a movie was released will play a significant role in whether the movie fails or passes the Bechdel test. More specifically, we predict that movies released in the later decades of our model will have a higher probability of passing the Bechdel test than those from earlier decades.

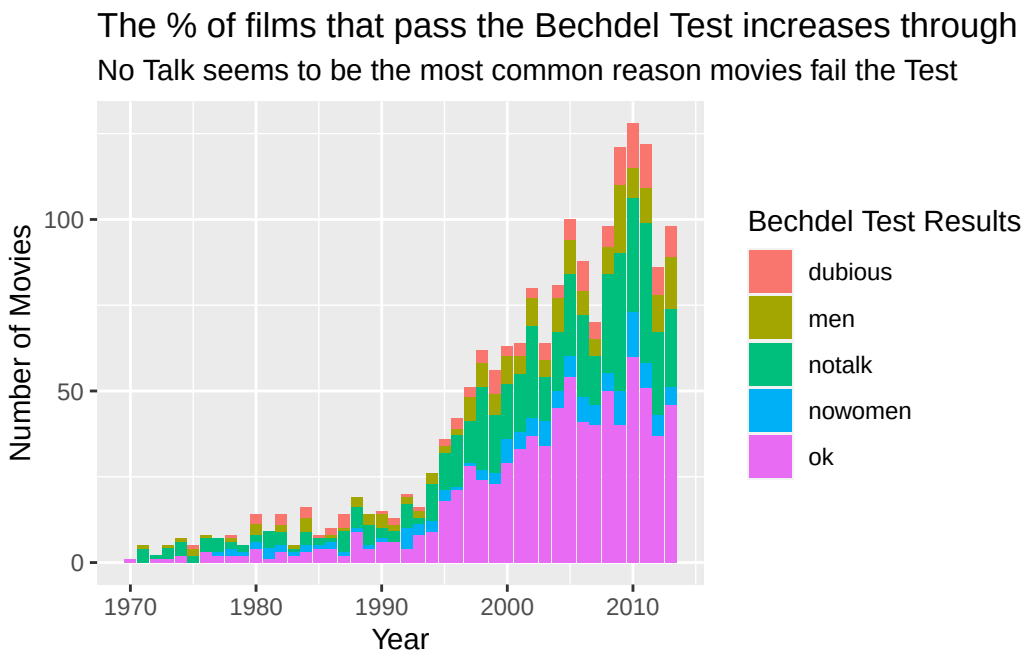
Data: To explore our research question, we'll be using a data set used in the FiveThirtyEight story titled "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women" (Hickey, 2014). The data set includes observations from 1794 films that were released between 1970-2013.The data set was organized by combiningdata from two major sources. One of them came

from the BechdelTest.com: a website operated by committed moviegoers who analyze films and ascertain if they pass the Bechdel test.To provide financial information for the chosen films, the FiveThirtyEight team gathered data from the website The-Numbers.com, a leading site for box office and budget data. The finalized *movies.cv* data set includes information detailing the title of the film, the year it was released, its domestic gross, budget and international gross (both accounting for inflation at the time of data collection and without). See Data Dictionary for more details. Furthermore, for our purposes, the data set also includes two important columns regarding whether the movie's performance on the Bechdel test. The column "binary" specifically states whether the movie passed or failed the Bechdel test in a binary fashion. The column "clean_test" goes a bit more in detail, regarding how the film failed the Bechdel test or if it was unclear whether the movie passed or not. The clean test variable has five levels: ok (Passed), no women (No women in the film), dubious (unclear result), no talk(women did not talk to each other), and men (women only talked about men). We created new variables for the purposes of our analysis. For starters, we created a decades variable that detailed which decade the given film was released in. The variable ended up having 5 different levels: 1970s, 1980s, 1990s, 2000s and 2010s. We also created a new variable titled "passfail" which is essentially the same as our 'binary' variable but instead uses dummy values to illustrated whether the movies passed (passfail=1) failed (passfail=0) the Bechdel test. We excluded 18 observations from our final analysis. These observations had no values for their total domestic gross and total budget. These are variables of interest that we want to control for. Since we still had 1776 observations left in our data set, our statistical analyses will not be greatly affected by the removal of these observations.

Exploratory Data Analysis



We can see above that there is general upward trend across the decades with an increased percentage of movies that pass the Bechdel test. We see, specifically that the movies from our data set that premiered in the 2000s, about 48% of the films passed the Bechdel test. The 2010s performed in a similar fashion, with 45% of the films released passing the Bechdel test (it is important to note however that this data set only includes movies till 2013, thus it does not paint the full picture of female representation in film from this decade) The decade with the lowest percentage of movies that passed the Bechdel test was the 1970s, withonly 25% passing. Below we also see a similar trend when looking years as our independent variable and with clean test as our dependent variable, revealing that films released in the 21st century have a higher percentage of movies that pass the Bechdel test. Moreover, we also see that "no talk" talk seems to be the most common reason for a movie to fail the Bechdel test.

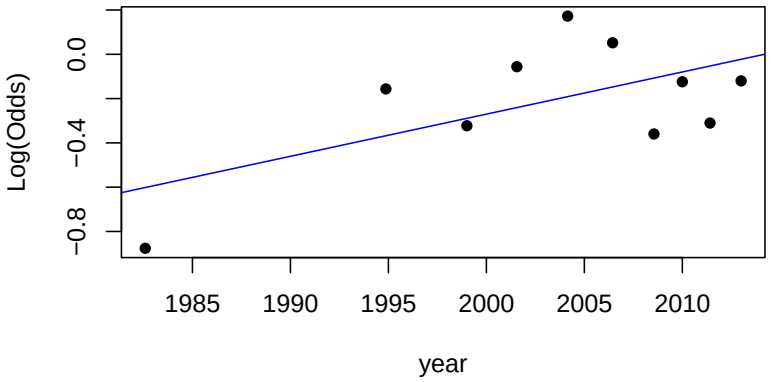


Methodolody

We are interested in running a logistic regression model in order to evaluate whether the time period a given film was released in is associated with its scoreon the Bechdel test. We believe this is an appropriate model, firstly, because our data passes the independence assumption. We can assume that independence is met because our observations are most likely not correlated with each other. Each of our movie titles are independent from each other and knowing something about one of our observations does not reveal anything substantial about another observation. For our outcome variable, we chose to use the binary version detailing whether given film's performance on the Bechdel test rather than our "cleantest" variable because we

felt that a binary outcome variable would generate a simpler model. Furthermore, if we ran a logistic regression model with the "cleantest" variable as our outcome variable, we would have to use a multinomial regression model to test out our research question. A multinomial regression model would not make sense in this context since the independence of irrelevant alternatives assumption would be violated. This assumption assumes that, in a multinomial logistic regression model, the relative odds of choosing one option over another should not be influenced by the inclusion or exclusion of an additional option. This does not make sense since the inclusion or exclusion of a Bechdel test result category could have an effect on our final analysis. For example, if a given film with plentiful female representation that was released in 2013 (which according to our hypothesis means it has a greater chance of passing the Bechdel test) was included our model but the only two categories taken accounted for whether "notalk" and "dubious," our model would predict it was it fit the 'dubious' category. However, if the "ok" was included in the mix, this would change our predictive probability.

Next, we evaluated whether to use the "year" or "decade" variable for our investigation. To decide which variable to use, we checked whether our continuous variable "year" passed the linearity assumption.



As shown above, the points are not evenly scattered through our line of best fit, therefore we have decided that this is not an appropriate variable to use in our model and instead will use the decade variable. We could have transformed the "year" variable, quadratically for example in order to pass this linearity assumption. However, this will complicate our interpretation of our model gravely since our outcome variable of interest would be transformed. Thus, for the purposes of our investigation, we will use the 'decades' categorical variable. For our other predictor variables,we are interested in using international gross, domestic gross, budget and an interaction between international gross and domestic gross. These variables have been shown to affect Bechdel test results in the past (Hickey, 2014) thus we thought it would be smart to control for these variables in our model.Furthermore, we chose to introduce an interaction term

between domestic gross and international gross because these two variables tend to depend on each other–when a movie does well internationally this depends on its domestic success since movie industries usually stop investing in marketing for the film if the film performs poorly domestically first since the film could be seen as a "lost cause" (stephenfollows.com).

We also checked to see if the linearity condition was met for our other continuous predictors (see Appendix for linearity plots for all continuous predictors). Only the budget variable passed the linearity assumption. The linearity condition was not met for our predictors dom_Gross2013 and Int_Gross2013. In order to deal with this violation of our linearity assumption, we applied a log transformation to these variables.

This is our proposed model that we will use to investigate our research question:

Our model is as follows. For each decade the film premiered in i,

$p/(1-p)$ = Odds of passing the Bechdel Test

$log(p/(1-p) = \beta_0 + \beta_1(Budget)_i + \beta_2(log(IntGross))_i + \beta_3(log(DomGross) + \beta_4 I(decade1980s)_i + \beta_5 I(decade1990s)_i + \beta_6 I(decade2000s)_i + \beta_7 I(decade2010s)_i + \beta_8(log(IntGross))_i * log(DomGross)_i$

Results

Hypothesis Test We will utilize logistic regression model using the final model discussed in the previous section to see if the decade in which a movie was released predicts whether a movie passes or fails the Bechdel test, while controlling for other variables in our model. We will first run a hypothesis test to see if there sufficient statistical evidence to suggest that there is a relationship between the decade in which a movie is released and whether the given film passes the Bechdel test.

We will run a hypothesis at the $a = 0.05$ level.

Null hypothesis: There is not sufficient evidence to suggest that the decade in which a movie premiered is associated with the differential odds of the passing the Bechdel test, while controlling for all of the variables represented in our model.

H_0 : All of our β terms for decade (β_4, β_5, β_6 and β_7) are equal to zero.

Alternative Hypothesis: There is sufficient evidence to suggest that the decade in which a movie premiered is associated with the differential odds of passing the Bechdel test, while controlling for all of the variables represented in our model.

H_1: At least one of our β terms for decade (β_4, β_5, β_6 or β_7) is not equal to zero.

```
# A tibble: 4 x 6
  term                                      Resid~1 Resid~2    df Devia~3  p.value
  <chr>                                       <dbl>   <dbl> <dbl>   <dbl>    <dbl>
1 passfail ~ decade + budget_Gross2013 +~    1767   2367.    NA    NA   NA
2 log(int_Gross2013) * log(dom_Gross2013)    1771   2397.    -4   -29.4  6.39e-6
3 passfail ~ budget_Gross2013 + log(dom_~    1767   2367.    NA    NA   NA
4 log(int_Gross2013) * log(dom_Gross2013)    1771   2397.    -4   -29.4  6.39e-6
# ... with abbreviated variable names 1: Resid..Df, 2: Resid..Dev, 3: Deviance
```

```
[1] 5.31495
```

We will be conducing an F test for this formal hypothesis test. The F statistic is 5.31495 and falls under an F distribution with 4 numerators of degrees of freedom and 1776 denominator degrees of freedom. We reject the null hypothesis in this case since our p value is less than 0.05, meaning that at least one of our decade predictor levels has a slope that is not 0. There is sufficient evidence to suggest that the decade in which a movie premiered in is associated with the differential odds of passing the Bechdel test, while controlling for all of the variables listed in the previous section (and adjusting for log(domGross), log(intgross) and log(intgross)* log(domGross.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0978128	2.5054732	-1.2364182	0.2163031
decade1980s	0.1186899	0.3735990	0.3176932	0.7507176
decade1990s	0.8667758	0.3391740	2.5555490	0.0106020
decade2000s	1.0851180	0.3314832	3.2735234	0.0010622
decade2010s	0.9558745	0.3392631	2.8175022	0.0048399
budget_Gross2013	0.0000000	0.0000000	-4.8974955	0.0000010
log(dom_Gross2013)	0.2098589	0.1789721	1.1725785	0.2409649
log(int_Gross2013)	0.0616945	0.1702408	0.3623954	0.7170566
log(dom_Gross2013):log(int_Gross2013)	-0.0077444	0.0095595	-0.8101225	0.4178698

Coefficient Interpretations

While controlling for the variables in our model, the odds of passing the Bechdel test for a movie that premiered in the 1990s is estimated to be 2.3379 times the odds of a film that was released in the 1970s (adjusting for our transformed variables). While controlling for the variables in our model, the odds of passing the Bechdel test for a movie that premiered in the 2000s is estimated to be 2.959 times the odds of a film that was released in the 1970s (adjusting for our transformed variables). While controlling for the variables in our model, the odds of passing the Bechdel test for a movie that premiered in the 2010s is estimated to be 2.601 times the odds of a film that was released in the 1970s (adjusting for our transformed variables).

Predictive Power: Even though our result was significant, we also want to check for the predictive power of our model. We will thus check the obtain predicted probabilities of success for our binary variable "passfail." We will impose a threshold of 0.5 as our classifier. We will check the sensibility and sensitivity for our model to asses its predictive power and construct an ROC curve to show how specificity and specificity change as our discrimination threshold changes (see ROC curve in Appendix). Calculating the area underneath the ROC curve will also give us more insight into the predictive power of our model.

.metric	.estimator	.estimate
roc_auc	binary	0.622274

Our Area under the curve is 0.62 which is higher than 0.05–meaning that our model suggests some discrimination its outcomes are not completely governed by chance.

	0	1
Fail	666	396
Pass	316	398

As shown by our table above, our specificity was 67.82%. Furthermore, our sensitivity was 50.13%.

Discussion

Ultimately, our model demonstrates that there is an association between the decade in which a film was released and whether it passes or fails the Bechdel test. This was illustrated through our hypothesis test which demonstrated a significant association between the decade in which a film was released and its score on the Bechdel test. Furthermore, our logistic regression model demonstrated that if a movie premiered during the 2010s, 2000s or the 1990s, it had higher odds of passing the Bechdel test compared to a movie that came out in the 1970s.

We can conclude, therefore, that the time period in which a movie came out in does in fact hold predictive power in predicting whether the given film passed or failed the Bechdel test. Perhaps the movie industry has indeed progressed throughout time and has released films with more adequate female representation in film. This could be a result of the recent diversity campaigns in Hollywood and a result of an increase of women working in the film industry, allowing for more adequate representation for female characters. However, the predictive power of our model was, overall, pretty weak. Our model only detected 50% of the films that in actuality passed the Bechdel test as shown by our sensitivity calculation. Our model did a slightly better job at predicting which movies failed the Bechdel test as shown by the fact that our model predicted 68% of films that in actuality failed the Bechdel test.

Our model most likely did not obtain enough predictive power because of the limitations our investigation faced. Firstly, our model had an uneven amount of films that premiered in our decades of interest. We had 54 observations for the 1970s, 125 observations for 1980s, 337 for 1990s, 840 for the 2000s and 438 for the 2010s (see Appendix for summary statistics).

This could have, in turn, affected our statistical inference since there was an unbalance of observations for each level of our predictor variables. Another limitation is the fact some of the coefficients in our model are pretty difficult to interpret due to our inclusion of transformed variables.Lastly, another limitation includes the lack of variables in our data set. Our data set only included 15 variables to begin with which have us a limited amount of variables to choose from and to control for in our final model. There are many future directions that could be taken with this investigation. For starters, it could be interesting to have a dataset that represents moredecades (from the 1920s till 2020s). Furthermore, future investigationsshould include more variables in their model. It could be interestingto see if the number of women working on the film is associated withthe film's score on the Bechdel test. Film genre could also be an interesting variable to look at; it is possible that movies from certain genres are predicted to have higher odds of passing the Bechdel test than others.

Sources

https://socialsciences.ucla.edu/wp-content/uploads/2022/03/UCLA-Hollywood- Diversity-Report-2022-Film-3-24-2022.pdf

https://www.nytimes.com/2017/02/21/movies/women-protagonists-movies-2016.html
https://www.tandfonline.com/doi/pdf/10.1080/14680777.2016.1234239? needAccess=true

https://bechdeltest.com/

https://fivethirtyeight.com/features/the-dollar-and-cents-case-against- hollywoods-exclusion-of-women/

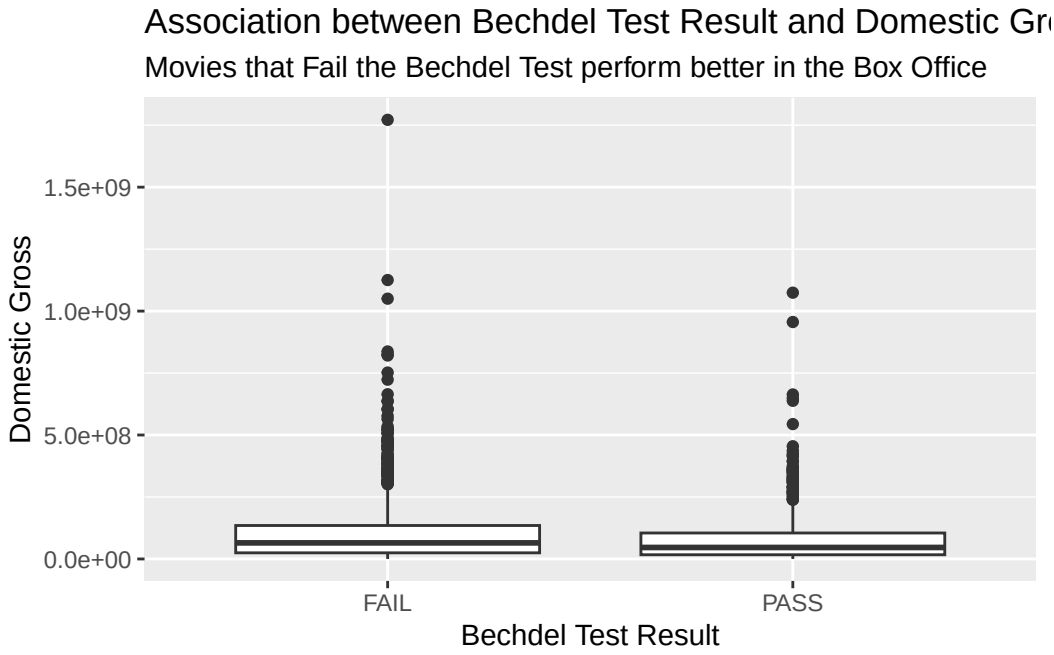
https://ew.com/oscars/2017/07/03/oscars-2017-new-members-diversity-push/

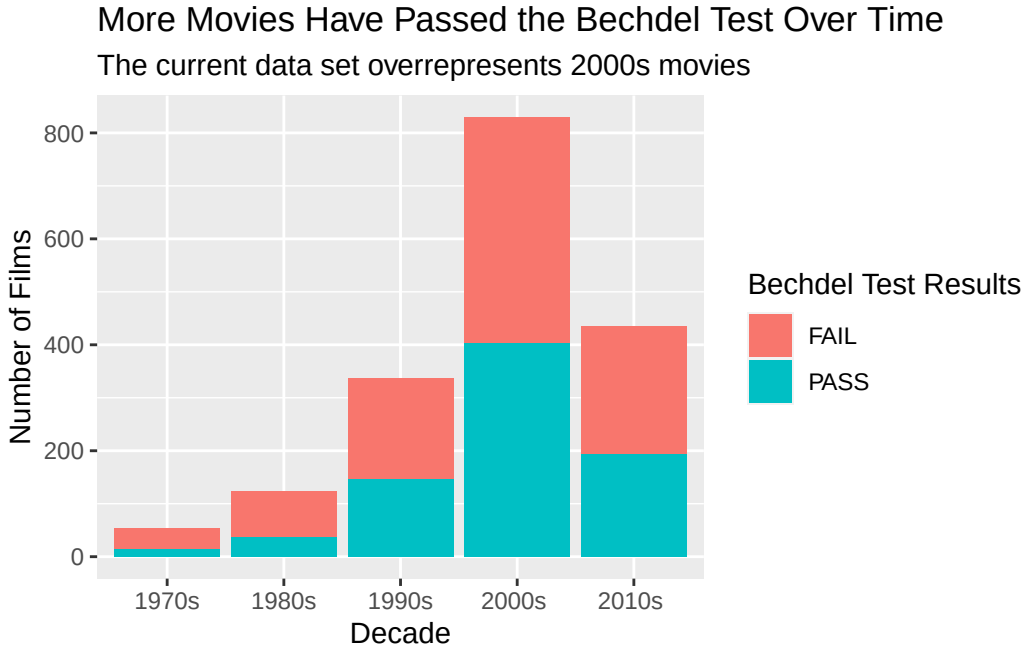
https://github.com/fivethirtyeight/data/tree/master/bechdel

https://stephenfollows.com/important-international-box-office-hollywood/

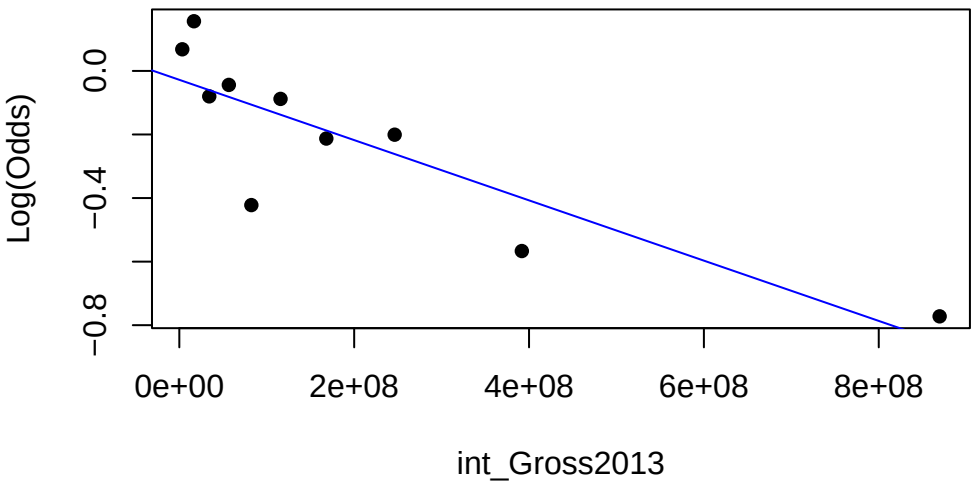
Appendix

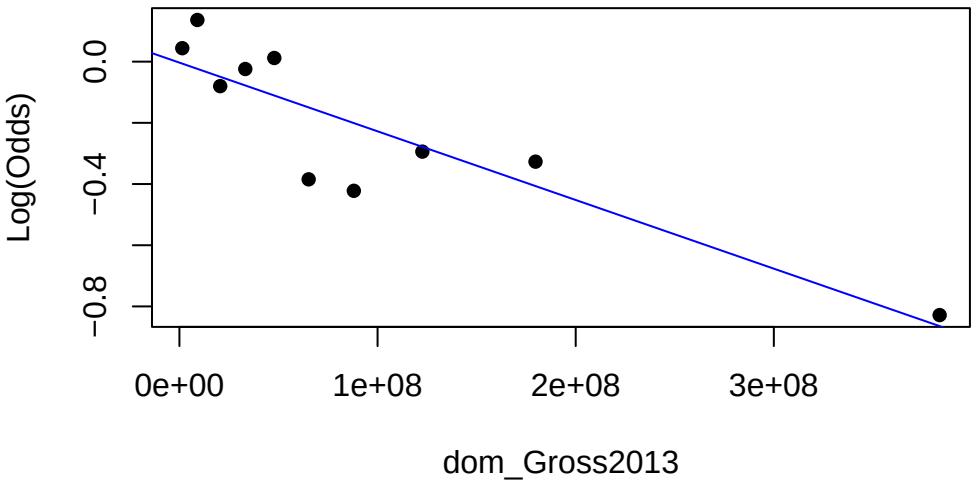
Additional Exploratory Analysis

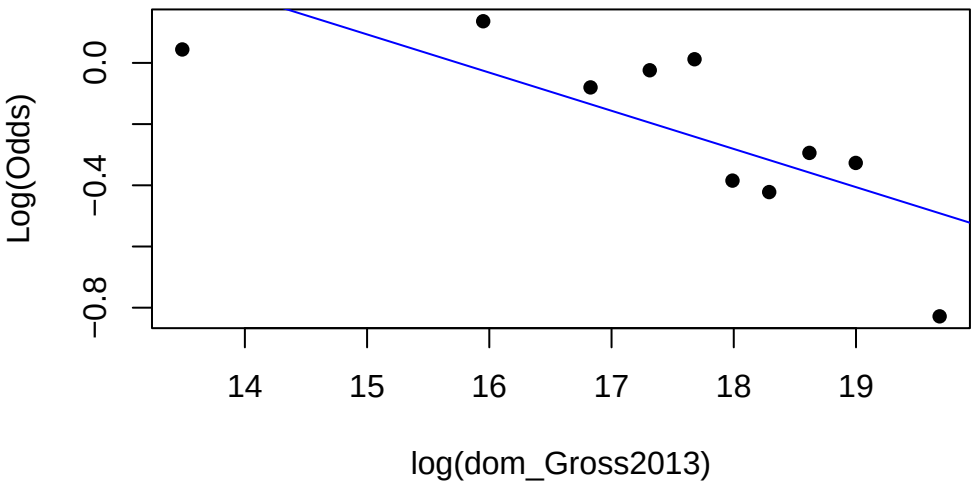


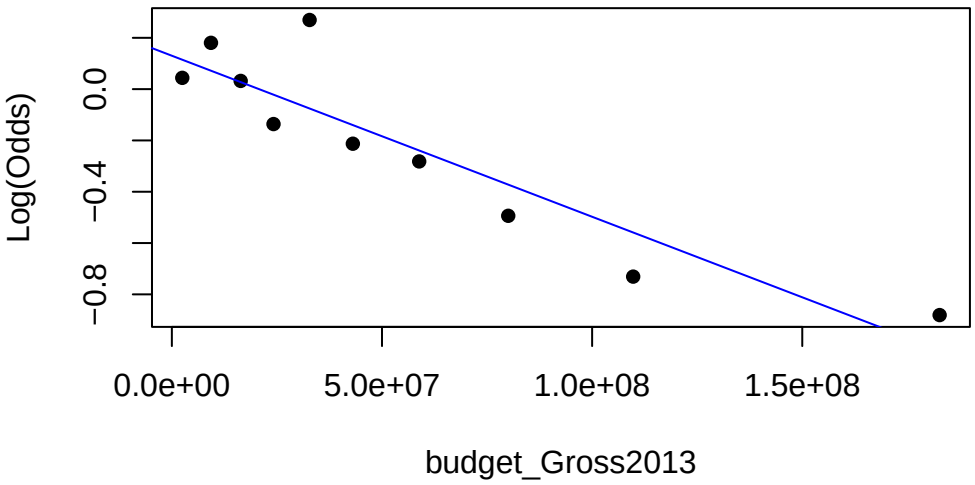


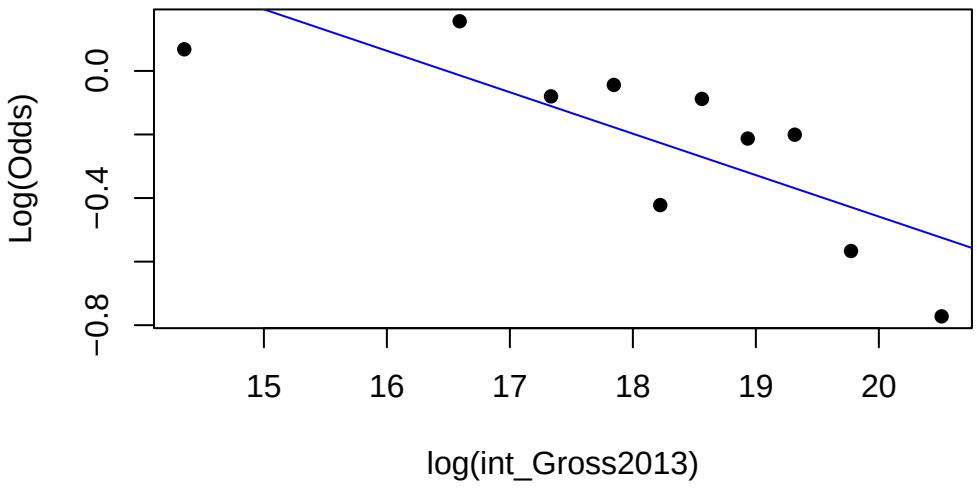
Linearity Plots for Continous Predictors

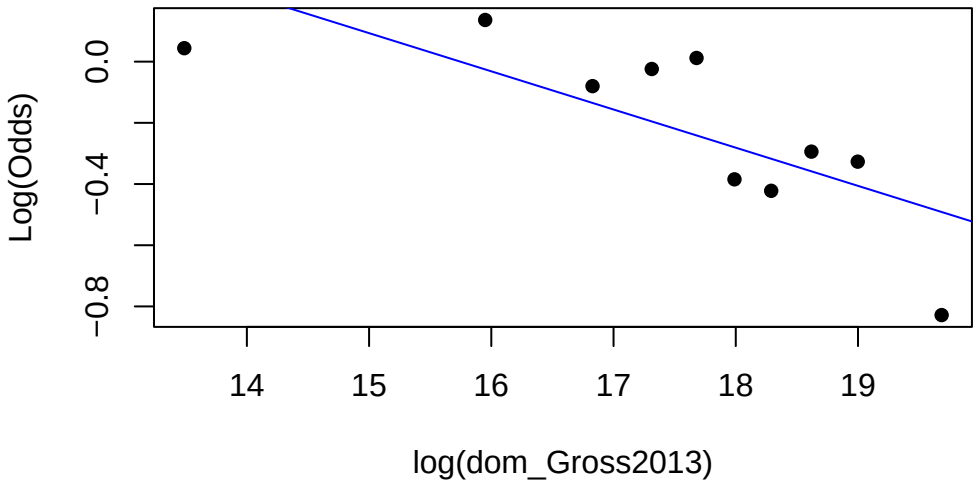

Log(Odds)
0.0
−0.4
−0.8
0e+00
2e+08
4e+08
6e+08
8e+08
int_Gross2013


Log(Odds)
0.0
−0.4
−0.8
0e+00
1e+08
2e+08
3e+08
dom_Gross2013

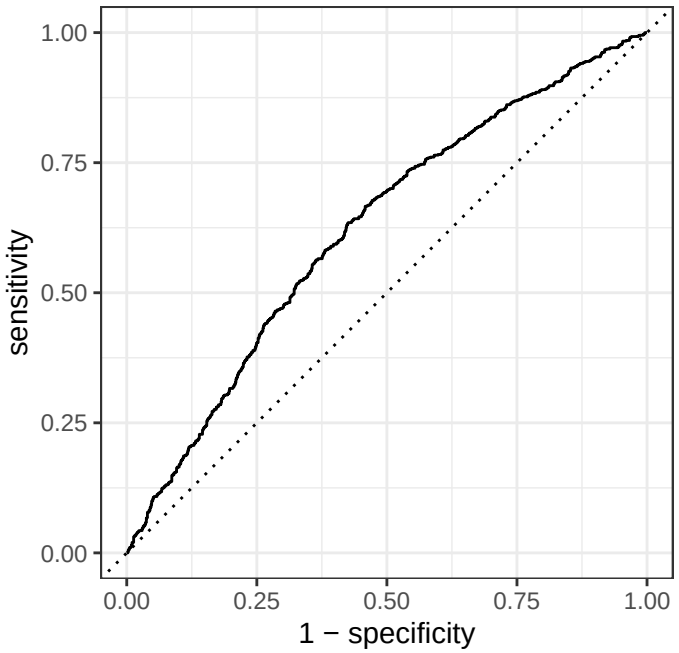

Log(Odds)
0.0
−0.4
−0.8
14
15
16
17
18
19
log(dom_Gross2013)


Log(Odds)
0.0
−0.4
−0.8
0.0e+00
5.0e+07
1.0e+08
1.5e+08
budget_Gross2013


Log(Odds)
0.0
−0.4
−0.8
15
16
17
18
19
20
log(int_Gross2013)


Log(Odds)
0.0
−0.4
−0.8
14
15
16
17
18
19
log(dom_Gross2013)

ROC Curve


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity