Stat 210 Project

Sofia Silvosa

Introduction

Background and Significance:

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). But there has been recent progress in the recent years with an impressive increase in female lead movies in the 2010s; not only that but these films have achieved monetary success. In 2016 with, 29 percent of protagonists in the top 100 box-office hits were 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 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). If this the case, then we would expect more recent films, specially those of the 2000s and 2010s, to have a greater amount of films that pass the Bechdel test.

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 specially, we predict that movies released in the 2000's and 2010's have a greater percentage of movies that pass the Bechdel rather than those that came out in the 20th century represented in our data set.

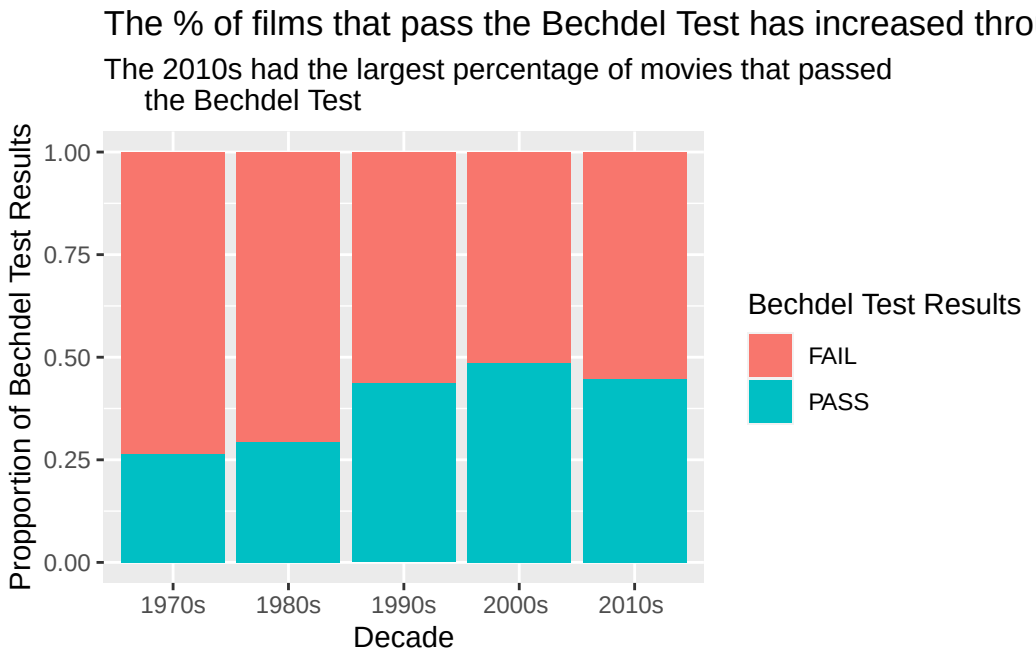
Data

To explore our research quetsion, 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 combining data 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.

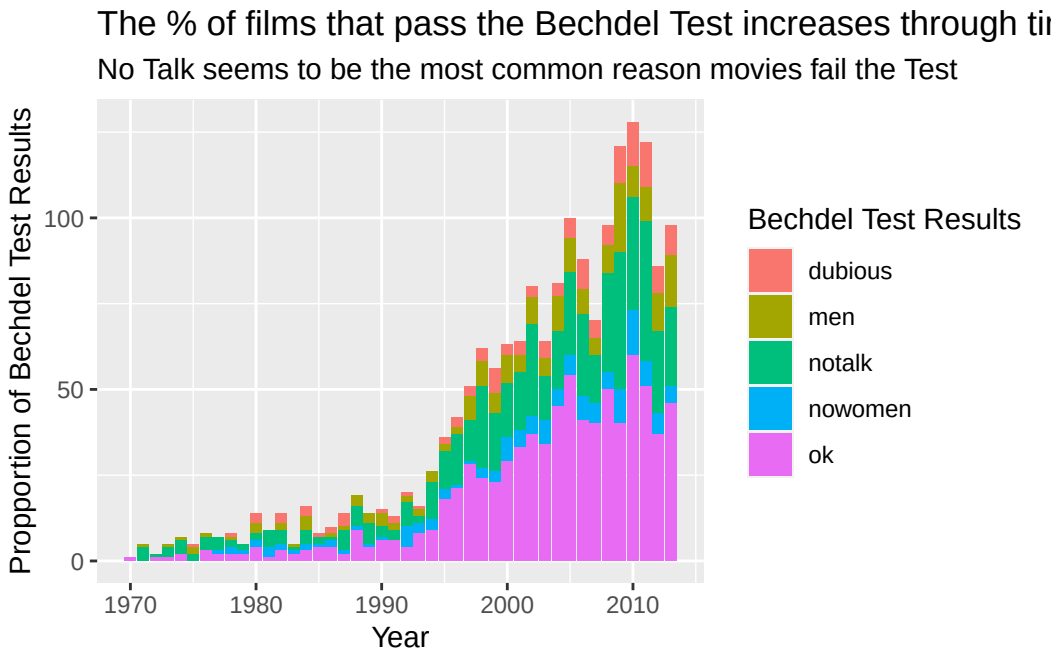
Furthermore, for our purposes, the data set also includes two important columns regarding whether the movie passed or failed the Bechdel test. The column "binary" specifically states whether the movie passed or failed the Bechdel test in a binary fashion. The column "clean_test" goes a bit more in detail, 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 important variables that we want to control for .Since we still had 1776 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, with only 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 women talk seems to be the most common reason for a movie to fail the Bechdel test.



#Methodolody

We are interested in running a regression model in order to evaluate whether the time period a given film was released. Furthermore, we are interested in using the binary variable "passfail" as our outcome variable–meaning we will use a logistic regression model. Furthermore, we believe this is an appropriate model since 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.

We chose to use the binary version detailing whether given film's performance on the Bechdel test rather than our clean_test variable because we felt that a binary outcome variable would generate a simpler model. Furthermore, if we ran a logistic model with the clean_test 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 the 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 failing 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 probablity.

Next, we evaluated whether to use the "year" or "decade" variable for our investigation. On one hand, the year variable is continouus, allowing us to have greater statistical power. On the other hand, the categorical "decade" variable applies more to our question of interest and the context of the data set. The film industry has gone through distinct decade period; many film research that tackles chronically change discuss the changes through decades, not years.

To decide which variable to use, we first calculated the RMSE for the variables.

Model 1:

$PassFail = \beta_0 + \beta_1(1980s)_i + \beta_2(1990s)_i + \beta_2(2000s) + \beta_3 I(2010s)$

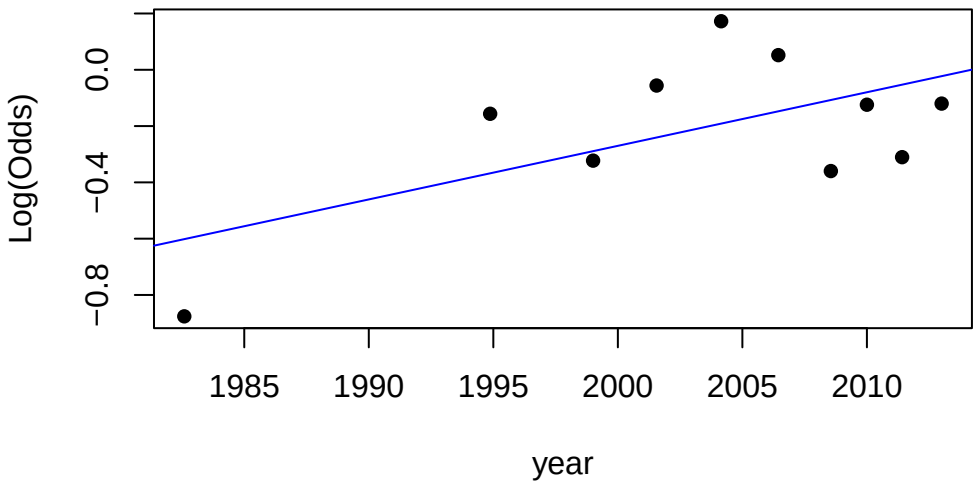
Model 2:

$PassFail = _0 + 1(year) + $

We calculate the Root Mean Squared Errors for the two models to see which predictor variable would be the smartest to use:

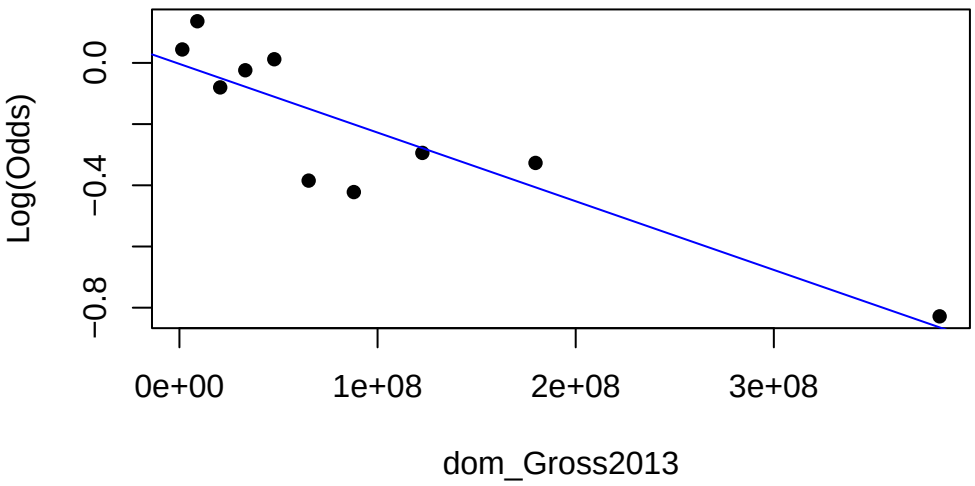
We found that our first model had a RMSE for our first model was 0.8505231 and for our second model the RMSE was 0.8373541. Thus, we decided to use the second model for our analysis.

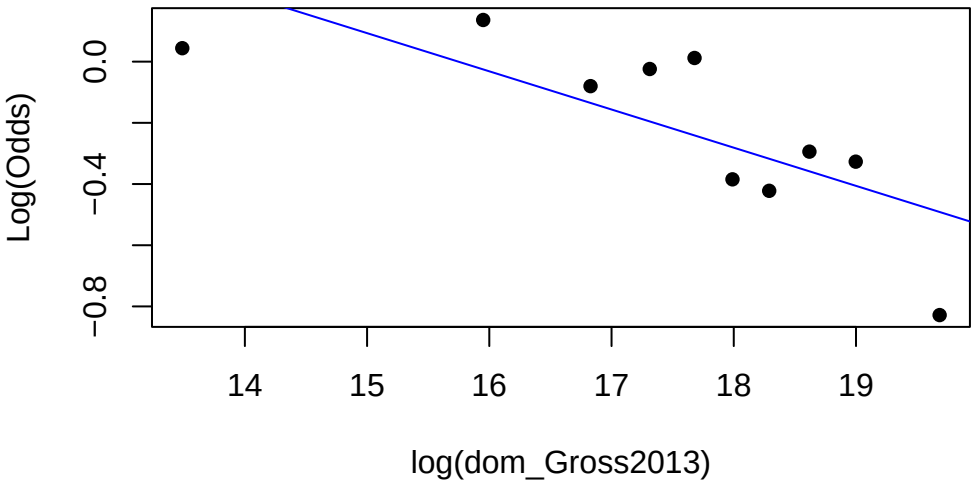
However, since we are using a logistic model, we are interested in seeing if it meets the logistic regression assumptions. And since year is a continous variable we decided to see if it met our linearity condition.

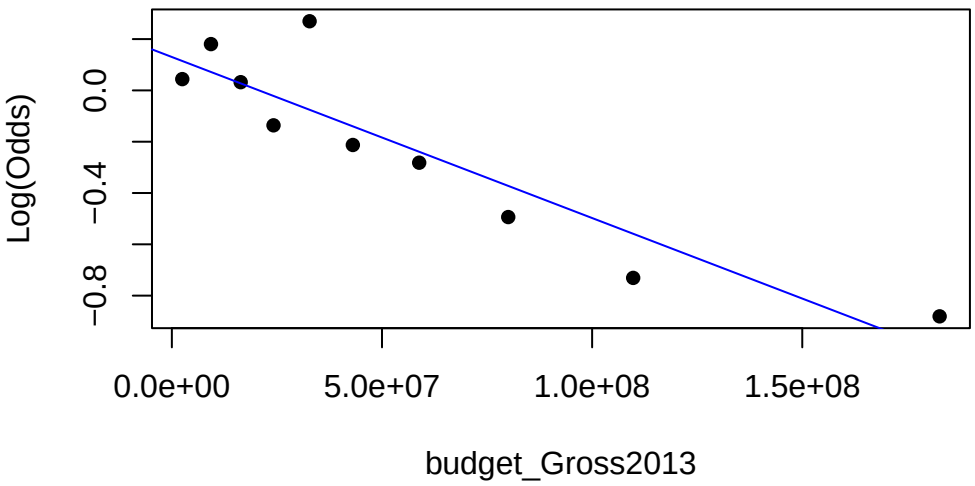


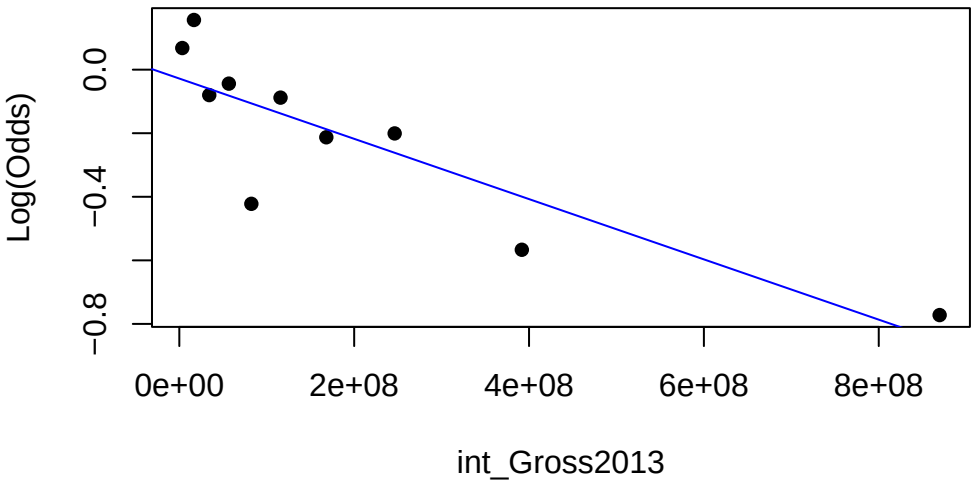
Points are not evenly scattered therefore we have decided that this is not an appropriate variable to use in our model and instead will use decade. We could have transformed the variable, quadratically for example in order to pass this linearity assumption. However, this will complicate our interpretation 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 variable.

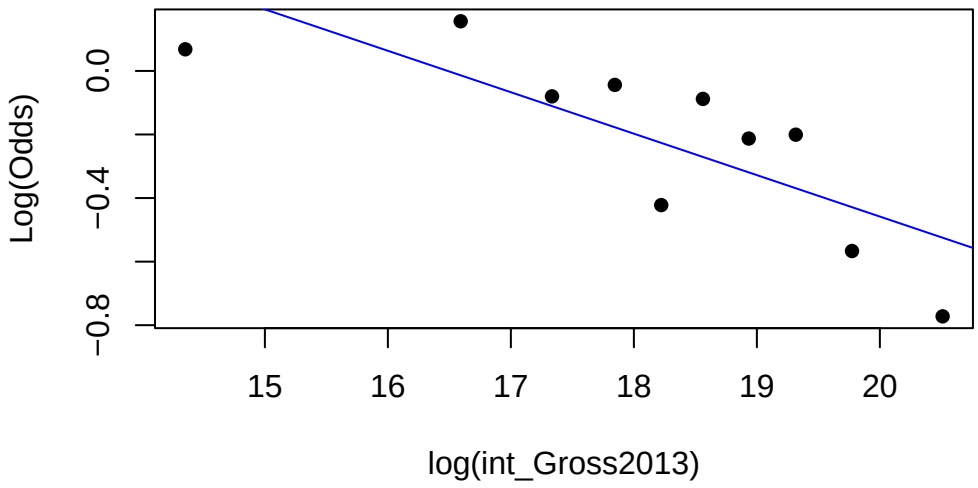
We also checked to see if the linearity condition was met for our other continuous predictors. Only the budget variable passed the linearity assumption.


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
0e+00
2e+08
4e+08
6e+08
8e+08
int_Gross2013



The linearity condition was met for our predictors dom_Gross2013 and budget_Gross2013. However, it was not met for our Int_Gross2013 variable. In order to deal with this violation of our linearity assumption, we applied a log transformation to our int_Gross variable. Since, this is not a particular variable of high interest, our final result interpretations' simplicity won't be gravely affected.

Furthermore, to further evaluate which variables to include in our data set, we ran a LASSO to determine if our variables of interest should be included in our model. They were all included thus we decided to include them in our model. We also decided to introduce an interaction term of intGross2013 * domGrpss2013 since in real life they depend on each other.

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

For each deccade the film premiered in i,

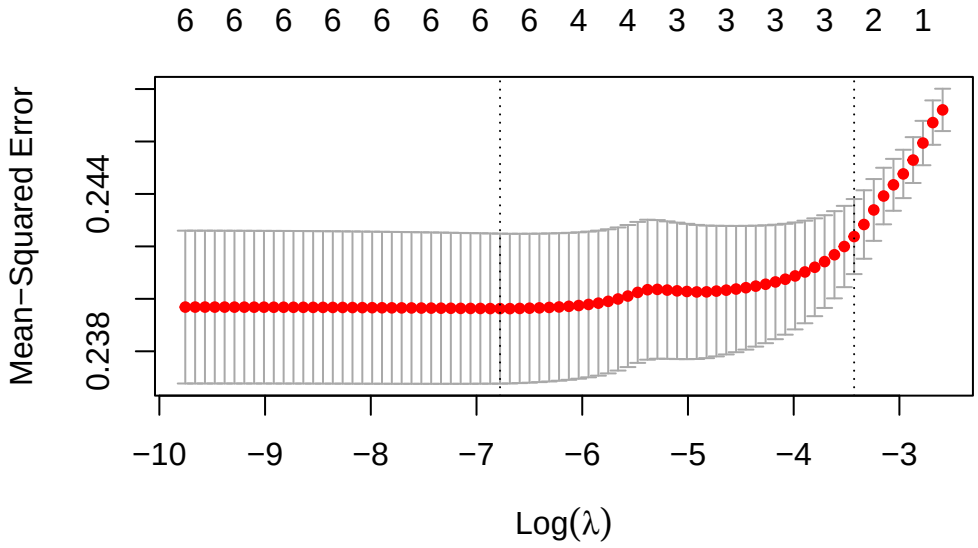
$p/(1-p)$ = Odds of passing the Bechdel Test

$log(p / (1-p) = _0 + _1(Budget)_i + _2(log(IntGross))_i + _3(log(DomGross)) + _4I(decade1980s)_i + _5I(decade1990s)_i + _6I(decade2000s)_i + _7I(decade2010s)_i + _8(log(IntGross))_i * log(DomGross_i) $

```
[1] 0.001138912
```

```
7 x 1 sparse Matrix of class "dgCMatrix"
                                        s0
```

```
(Intercept)                    .
year                           4.192155e-03
budget_Gross2013              -1.624111e-09
dom_Gross2013                 -6.183545e-10
log(int_Gross2013)             2.148681e-03
int_Gross2013                  2.703017e-10
dom_Gross2013:int_Gross2013    1.235425e-20
```



Results

Hypothesis Test We ran a logistic regression model using the final model discussed in the previous section to see if decade alone can predict whether a movie passes or fails the Bechdel test. We will then run a hypothesis test to see if one of our decade categorical predictors is associated with the score a film gets on the Bechdel test.

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

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

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

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

Null hypothesis:

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

There is not sufficient evidence to suggest that the decade in which a movie premiered is associated with differential odds of the moving passing the Bechdel test, while controlling for all of the variables listed in the previous section.

Alternative Hypothesis:

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 +~    1765   2357.    NA    NA   NA
2 int_Gross2013 * log(dom_Gross2013)         1769   2379.    -4   -21.1  2.96e-4
3 passfail ~ budget_Gross2013 + dom_Gros~    1765   2357.    NA    NA   NA
4 int_Gross2013 * log(dom_Gross2013)         1769   2379.    -4   -21.1  2.96e-4
# ... 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 freeom 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 moviepremiered in is associated with differential odds of the moving passing the Bechdel test, while controlling for all of the variables listed in the previous section (andadjusting for log(domGross), log(intgross) and log(intgross)* log(domGross.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2260560	0.9099559	-1.3473796	0.1778580
decade1980s	-0.0304507	0.3840440	-0.0792895	0.9368023
decade1990s	0.6285726	0.3566376	1.7624968	0.0779854
decade2000s	0.8578170	0.3500244	2.4507348	0.0142565
decade2010s	0.6952480	0.3603416	1.9294135	0.0536795
budget_Gross2013	0.0000000	0.0000000	-5.8222918	0.0000000
dom_Gross2013	0.0000000	0.0000000	-2.9262219	0.0034311
log(int_Gross2013)	-0.1302922	0.1030034	-1.2649306	0.2058962
int_Gross2013	0.0000000	0.0000000	-0.0269941	0.9784644
log(dom_Gross2013)	0.1820920	0.0904959	2.0121578	0.0442033
int_Gross2013:log(dom_Gross2013)	0.0000000	0.0000000	0.2923768	0.7699985

```
[1] 2.358008
```

```
[1] 2.004206
```

```
Call:
glm(formula = passfail ~ decade + budget_Gross2013 + log(dom_Gross2013) +
    log(int_Gross2013) + log(int_Gross2013) * log(dom_Gross2013),
    family = "binomial", data = movies1_0)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3355  -1.1236  -0.8081   1.1603   2.0203

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -3.098e+00  2.505e+00  -1.236  0.21630
decade1980s                           1.187e-01  3.736e-01   0.318  0.75072
decade1990s                           8.668e-01  3.392e-01   2.556  0.01060
decade2000s                           1.085e+00  3.315e-01   3.274  0.00106
decade2010s                           9.559e-01  3.393e-01   2.818  0.00484
budget_Gross2013                     -6.323e-09  1.291e-09  -4.897 9.71e-07
log(dom_Gross2013)                    2.099e-01  1.790e-01   1.173  0.24096
log(int_Gross2013)                    6.169e-02  1.702e-01   0.362  0.71706
log(dom_Gross2013):log(int_Gross2013) -7.744e-03  9.560e-03  -0.810  0.41787

(Intercept)
decade1980s
decade1990s                          *
decade2000s                          **
decade2010s                          **
budget_Gross2013                     ***
log(dom_Gross2013)
log(int_Gross2013)
log(dom_Gross2013):log(int_Gross2013)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2442.1  on 1775  degrees of freedom
Residual deviance: 2367.2  on 1767  degrees of freedom
```